جامعة الشاذلي بن جديد - الطارف
Chadli Bendjedid El Tarf University

CHADLI BENDJEDID UNIVERSITY

كلية العلوم و التكنولوجيا
FACULTY OF SCIENCE AND TECHNOLOGY

# THE 2ND INTERNATIONAL CONFERENCE ON AUTONOMOUS SYSTEMS AND THEIR APPLICATIONS

قسم الإعلام الآلي
INFO
DÉPARTEMENT D'INFORMATIQUE

# ICASA'24

HOUSE OF ARTIFICIAL INTELLIGENCE
CHADLI BENDJEDID UNIVERSITY - EL TARF

EDITED BY

ALI ABDELLATIF BETOUIL

ABDELMADJID BENMACHICHE

CHAOUKI CHEMAM

KHADIJA RAIS

MAY 2024

ICASA 2024 Conference Presentations

| **Presentation Title** | **Presenter(s)** |
|---|---|
| A CONVOLUTIONAL NEURAL NETWORK FOR SLEEP APNEA DETECTION | **Amal Miloud Aouidate** |
| A Fuzzy-KNN approach for Heart disease Prediction | **Amal Miloud Aouidate** |
| A Machine Learning Algorithms for decision-makers and analysts in the field of digital finance | **Atmane Hadji** |
| A system for detecting traffic objects and estimating their distance | **Bouguerne Innen** |
| Adaptive Shortest Path Algorithms in Dynamic and Uncertain Environments | **Boutabia Ines** |
| Advancements in Suicide Ideation Detection: A Comprehensive Literature Review | **Ferdadus BENROUBA** |
| AI-Driven Cybersecurity Orchestration and Automation | **Lokman TOUIL** |
| Applications of Secure Multi-Party Computation in Financial Services | **Sedraoui Brahim Khalil** |
| Collaborative Filtering Models Analysis on Amazon Products Dataset | **Maroua BENLEULMI** |
| Diagnosis of Induction Machine and Severity Estimation using Novel Gradient Boosting technique | **Bentrad Moutaz Bellah** |
| Early Parkinson's Detection: A Speech-Based Deep Learning Model with LSTM for Accurate Diagnosis | **Djaidja Imane** |
| Enhancing Medical Image Analysis through Geometric and Photometric Transformations | **Rais Khadija** |
| Enhancing Security in E-Learning Platforms | **Chennan Chaouki** |
| Enhance Trajectory prediction based swarm with deep learning | **Gahem Abderrahmen** |
| Ensuring Data Security and Privacy in Smart City Surveillance Systems Using Chaotic Maps and Smart Contracts | **Mohamed el amine KHERAIFIA** |
| Explainable AI in Decision-Making: Benefits and Challenges | **Nour El Houda Dehimi** |
| Exploring the Integration of Differential Privacy in Cybersecurity Analytics: Balancing Data Utility and Privacy in Threat Intelligence | **Sedraoui Brahim Khalil** |
| Hybrid Multi-Factor Authentication (MFA) Using Biometrics and Behavioral Analysis | **BOUFAIDA Soundes Oumaima** |
| Hybrid Secure Routing in Mobile Ad-hoc Networks (MANETSs) | **BOUFAIDA Soundes Oumaima** |
| Improving Recommendation quality via Ensemble Neural Networks | **Ramzi Khantouchi** |
| Improving Scalability in Block Chain based Federated Learning Using Off-chain Storage | **Ala Djeddai** |
| Intelligent Fault Detection base on machine learning in Autonomous Vehicles: A Hybrid Approach | **Lokman TOUIL** |

| Presentation Title | Presenter(s) |
|---|---|
| Intelligent Industrial Process Monitoring: SSAE for Quality Assurance | **Wafa Bougheloum** |
| New mapping method based on Bat algorithm for Embedded Systems | **Farid Boumaza** |
| Optimizing E-Commerce Product Recommendations Using Reinforcement Learning | **Abderaouf Bahi** |
| Optimizing Logistic Strategies in Algeria through Artificial Intelligence: A Comprehensive Analysis | **Besma AHMED MALEK** |
| Optimizing Urban Traffic Safety: A Case Study on Intelligent Transportation Systems in Algeria | **Nada Ahmed Malek** |
| Paving the Way for Smart Cyber Security: Integrating Artificial Intelligence and Building a Roadmap | **Zina Oudina** |
| SE-GNN: A Social-Enhanced Graph Neural Network for Personalized Recommendation | **Sara Gasmi** |
| Segmentation based deep learning approach for corrosion detection | **Abid Safa** |
| Sentiment Analysis of the Maghrebi Dialect Using RoBERT | **Faiz Maazouzi** |
| Shortest Path Algorithms for Autonomous Systems and Robotics | **Boutabia Ines** |
| Security Risk Modeling for Critical Systems: Attack Graphs and Attack Trees | **Zina Oudina** |
| Toward a Secured Communication Protocol for Unmanned Aerial Vehicles: A Security Map and Cryptography | **Zina Oudina** |
| Unmanned Aerial Vehicles Security: Threat Analysis and Hybrid Cryptography for secured Communication | **Zina Oudina** |
| Using of the A-OL Specification Language for the Automatic Generation of a Memory Controller Specifically for High Efficiency Video coding (HEVC) | **Messaoudi Newfel** |
| VerChain: Blockchain Based Certificate Degree Attestation and Verification in Algeria | **Rofaida Khemaissia** |

# A CONVOLUTIONAL NEURAL NETWORK FOR SLEEP APNEA DETECTION

Manare DINE [1], Amal MILOUD AOUIDATE [1]

[1] *Chadli Bendjedid University, El-Tarf, Algeria*

## Abstract

This work focuses on the development of an innovative model for sleep apnea detection based on analyzing electrocardiographic (ECG) signals. Sleep apnea, a condition characterized by repeated interruptions in breathing during sleep, poses severe health risks, including cardiovascular complications, hypertension, and an overall decline in quality of life. Timely and accurate detection of this condition is essential for effective management and the prevention of harmful consequences.

In this research, we leverage recent advancements in machine learning, particularly convolutional neural networks (CNN). We have created a combine CNN-KNN model that combines the data processing capabilities of CNNs with the simplicity and efficiency of the k-nearest neighbors (KNN) algorithm, aiming to enhance the accuracy of apneic event classification. The results indicate that our CNN-KNN approach significantly improves accuracy compared to using either CNN or KNN independently.

## Keywords

Sleep apnea, ECG, Apnea detection, Diagnosis, CNN, KNN, Deep learning, combination of models, Classification.

## 1. Introduction

The motivations behind the search for effective solutions to sleep apnea detection are both diverse and critical. First and foremost, there is a public health imperative, as sleep apnea is a prevalent yet frequently overlooked disorder that can lead to serious long-term health issues. Additionally, there is a strong desire to enhance the quality of life for those affected by this condition by providing more accessible and less invasive diagnostic tools.

This work aims to utilize artificial intelligence to develop a robust and effective sleep apnea detection model. Initially, we will focus on designing signal processing algorithms and machine learning models, particularly convolutional neural networks (CNNs) and K-Nearest Neighbors (KNN) classifiers, to analyze electrocardiogram (ECG) data and detect apnea signals.

## 2. Sleep Apnea Definition

Sleep Apnea Syndrome, also known as Obstructive Sleep Apnea Syndrome (OSAS), is a common condition characterized by frequent interruptions in breathing during sleep, lasting from 10 to 30 seconds and occurring at least five times per hour. This condition is often associated with pharyngeal collapse and airway obstruction, which can lead to hypoxia. In response, the brain prompts the individual to wake up in order to resume breathing. The condition is typically classified into three stages, ranging from mild to severe sleep apnea. [W2]

## 3. Severity Levels of Sleep Apnea

The Apnea-Hypopnea Index (AHI) quantifies the total number of apnea and hypopnea episodes relative to the total duration of sleep during which these events occur. The formula for calculating AHI is as follows [2]:

$$AHI = \frac{Number\ of\ apneas +\ Number\ of\ hypopneas}{Sleep\ duration\ (min)} \times\ 60$$

[1] Corresponding author.
These authors contributed equally.
dinemanare@gmail.com (M. Dine) ; a.miloudaouidate@univ-eltarf.dz (A. MiloudAouidate)

This index enables the classification of the severity of Sleep Apnea (SA) into three distinct levels [2]:

- Mild Sleep Apnea: AHI between 5 and 15.
- Moderate Sleep Apnea: AHI between 15 and 30.
- Severe Sleep Apnea: AHI greater than 30.

Given that apneas and hypopneas can have detrimental effects on the body, it is essential to be aware of these issues and to recognize the symptoms associated with this condition.[2]

## 4. Related Works

Research on sleep apnea detection using CNNs has focused on enhancing the reliability and efficiency of detection systems by exploring various methods of feature extraction and classification. Key findings from studies include:

- Tao Wang and Changhua utilized a modified LeNet-5 CNN for automatic feature extraction, surpassing traditional methods like SVM, KNN, LR, and MLP that relied on RR intervals and amplitude-based features. [5]
- Yunxiang Bai's research applied 1D convolution for intrinsic feature extraction from ECG signals, followed by fully connected layers for classification. Cross-entropy loss and hyperparameter tuning improved model performance. [7]
- A study introduced a Time Windowed Multi-Layer Perceptron (TW-MLP) to capture temporal dependencies in ECG segments, achieving superior results compared to traditional techniques. [4]
- Another work compared traditional methods (SVM, KNN, LR, MLP) with a novel approach using SE-ResNext 50 for automatic feature extraction. [3]
- Hung-Yu Chang's team employed a 1D CNN on raw ECG signals, bypassing complex preprocessing steps like QRS detection. With minimal preprocessing. [1]

## 5. Sleep Apnea ECG Dataset

Through the public research platform PhysioNet, which provides free access to large collections of biomedical and physiological data, we utilized the Apnea-ECG dataset, which provides a rich collection of 70 recordings, carefully divided into two distinct groups for comprehensive analysis [W1]:

- The first group, consisting of 35 recordings from a01 to a20, b01 to b05, and c01 to c10, serves as our training set.[6]
- The second group, also comprising 35 recordings from x01 to x35, represents our test set. [6]

It should be noted that each of the 70 recordings consists of a set of twelve files to process, resulting in a total of 840 files. [6]

## 6. The Proposed Methodology

In our sleep apnea detection model, we utilized two distinct methodological approaches: The K-Nearest Neighbors (KNN) classifier and a Convolutional Neural Network (CNN), along with a combined approach of CNN + KNN.

### 6.1. KNN Classifier Approach

For the KNN algorithm-based method, we adopted a systematic methodology. Initially, we preprocessed the data by normalizing the R-R intervals (RRI) and amplitude signals. To address potential class imbalance during model training, we employed the Random Over Sampler technique to ensure balanced classes. This step is critical for achieving robust model performance, particularly in scenarios with class imbalance.

After determining the optimal parameters, we trained the KNN model on the training dataset using these parameters. Upon completion of the training, we evaluated the model's classification performance on the test dataset, specifically assessing its ability to differentiate between the "Apnea" and "No-apnea" classes.

To gain a comprehensive understanding of the model’s performance, we calculated both precision and the confusion matrix.

## 6.2. Convolutional Neural Network

### 6.2.1. Choice of the CNN Model

To find the optimal model for data classification, we examined several CNN architectures. Each model was constructed with varying layers, hyper parameters, and regularization techniques. Our objective was to identify a model that strikes the right balance between generalization ability and performance on test data. We compared results based on various metrics, including accuracy and loss.

The tested models were as follows:

- Model 1: Conv1D-32-5-ReLU-MaxPool1D-2-Flatten-Dense-32-ReLU-Dense-2-Softmax
- Model 2: Conv1D-32-5-ReLU-MaxPool1D-3-Conv1D-64-5-ReLU-MaxPool1D-3-Flatten-Dropout-0.8-Dense-32-ReLU-Dense-2-Softmax
- Model 3: Conv1D-64-5-ReLU-MaxPool1D-2-Dropout-0.5-Conv1D-64-5-ReLU-MaxPool1D-2-Dropout-0.5-Concat-Conv1D-128-3-ReLU-MaxPool1D-2-Dropout-0.5-Flatten-Dense-64-ReLU-Dense-2-Softmax
- Model 4: Conv1D-64-5-ReLU-MaxPool1D-3-Conv1D-128-5-ReLU-MaxPool1D-3-Flatten-Dropout-0.5-Dense-64-ReLU-Dense-2-Softmax

Among these models, the fourth model emerged as the most effective, achieving an impressive accuracy of 89% and a minimal loss of 0.25 during evaluation. Its architecture consists of 1D convolutional layers followed by pooling layers, which effectively extract key features from the data. The incorporation of a 50% dropout layer helps to prevent overfitting, while the dense layers facilitate precise classification through ReLU activation. With its outstanding performance, this model demonstrates both robustness and generalization capabilities, making it an optimal choice for binary classification tasks.

### 6.2.2. CNN Model Architecture

The architecture depicted in the following figure illustrates the functional Convolutional Neural Network (CNN) developed for the classification of sleep apnea data.

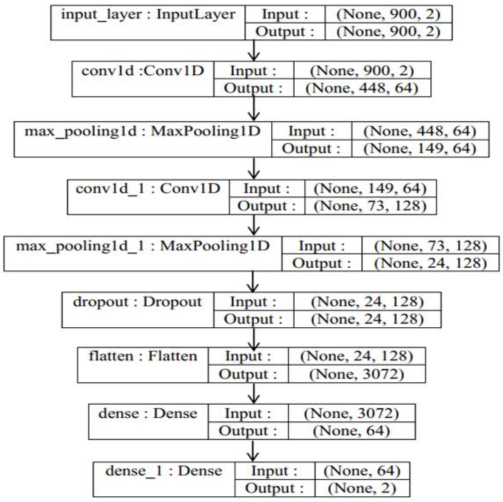
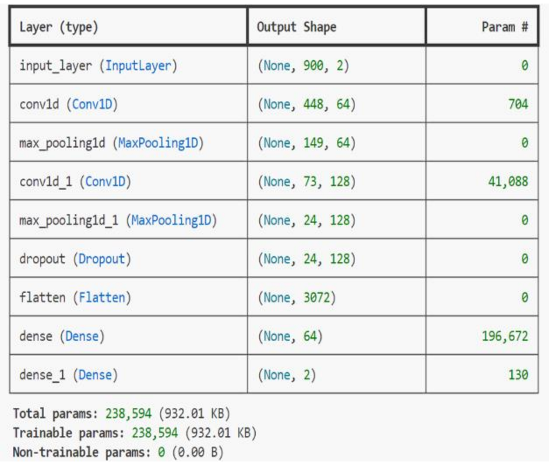

input_layer : InputLayer | Input : (None, 900, 2) | Output : (None, 900, 2)
conv1d :Conv1D | Input : (None, 900, 2) | Output : (None, 448, 64)
max_pooling1d : MaxPooling1D | Input : (None, 448, 64) | Output : (None, 149, 64)
conv1d_1 : Conv1D | Input : (None, 149, 64) | Output : (None, 73, 128)
max_pooling1d_1 : MaxPooling1D | Input : (None, 73, 128) | Output : (None, 24, 128)
dropout : Dropout | Input : (None, 24, 128) | Output : (None, 24, 128)
flatten : Flatten | Input : (None, 24, 128) | Output : (None, 3072)
dense : Dense | Input : (None, 3072) | Output : (None, 64)
dense_1 : Dense | Input : (None, 64) | Output : (None, 2)

**Figure 1:** Sleep Apnea CNN Model Architecture.

Here is the configuration and parameter settings of the developed AS model:



```
┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                     ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ input_layer (InputLayer)         │ (None, 900, 2)         │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ conv1d (Conv1D)                  │ (None, 448, 64)        │           704 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ max_pooling1d (MaxPooling1D)     │ (None, 149, 64)        │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ conv1d_1 (Conv1D)                │ (None, 73, 128)        │        41,088 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ max_pooling1d_1 (MaxPooling1D)   │ (None, 24, 128)        │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ dropout (Dropout)                │ (None, 24, 128)        │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ flatten (Flatten)                │ (None, 3072)           │             0 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ dense (Dense)                    │ (None, 64)             │       196,672 │
├──────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                  │ (None, 2)              │           130 │
└──────────────────────────────────┴────────────────────────┴───────────────┘
 Total params: 238,594 (932.01 KB)
 Trainable params: 238,594 (932.01 KB)
 Non-trainable params: 0 (0.00 B)
```

**Figure 2:** Configuration and parameter settings of the Sleep apnea model.

## 7. Discussion and Evaluation of the Results Obtained

In this section, we will present and analyze the results obtained the investigated approaches: first, our custom architecture developed earlier, utilizing only a convolutional neural network (CNN); second, the application of the K-Nearest Neighbors (KNN) classification algorithm independently; and finally, the evaluation of a combination of CNN and KNN methods.

### 7.1. Results and Evaluation of the AS

### 7.1.1. Model (CNN)

| | Apnea | No-Apnea | Macro-avg | Weighted-avg |
|---|---|---|---|---|
| Precision | 0.94 | 0.80 | 0.87 | 0.89 |
| Recall | 0.88 | 0.89 | 0.89 | 0.89 |
| F1-Score | 0.91 | 0.85 | 0.88 | 0.89 |

**Table 1:** Performance Evaluation of the Sleep Apnea Detection Model (CNN).

Table 1 illustrates the performance evolution of the sleep apnea detection model utilizing the CNN architecture, it evaluates the model's ability to accurately classify samples based on their class labels, distinguishing between apnea and non-apnea samples:

- Precision: Measures the proportion of correctly classified samples for each class. For "Apnea," precision is 0.94, meaning 94.22% of apnea predictions are correct. For "No-Apnea," it is 0.80, with 80.99% correctly identified.
- Recall: Indicates the proportion of actual samples correctly identified. The "Apnea" recall is 0.88 (88.90% accuracy), and "No-Apnea" recall is 0.89 (89.65% accuracy).
- F1-Score: Combines precision and recall into a single measure. The F1-score is 0.91 for "Apnea" and 0.85 for "No-Apnea."
- Macro-avg: Provides the unweighted average of metrics across classes. Precision and recall average 0.89, with an F1-score of 0.88, reflecting general performance.
- Weighted-avg: Calculates the weighted average, considering class proportions. Precision, recall, and F1-score are all 0.89, adjusting for class imbalance.

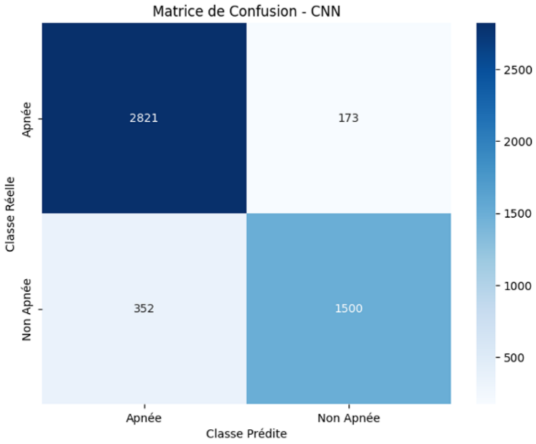


**Figure 3:** The Confusion Matrix of the CNN Model.

The confusion matrix presented in Figure 34 offers a comprehensive view of the performance of the binary classification model for sleep apnea detection. In this matrix:

- True Positives (TP): These represent the cases where the model correctly identified the positive class, totaling 2821 samples.
- True Negatives (TN): These are the cases where the model accurately predicted the negative class, amounting to 1500 samples.
- False Positives (FP): These refer to instances where the model incorrectly classified the positive class when it was actually negative, accounting for 352 samples.
- False Negatives (FN): These are the cases where the model incorrectly predicted the negative class when it was truly positive, totaling 173 samples.

### 7.1.2. Results and Evaluation of the CNN+KNN Model

This study employs a combination of two machine learning models: a Convolutional Neural Network (CNN) and a K-Nearest Neighbors (KNN) algorithm to classify sleep apnea using electrocardiography (ECG) data. The data is initially loaded from a preprocessed file in pickle format, followed by further preprocessing and splitting into training and test sets.

A CNN model is subsequently developed and trained on the apnea dataset, incorporating an optimization configuration along with an Early-Stopping callback to mitigate overfitting. The model's performance is assessed using the test data, from which the loss and accuracy of the CNN alone are calculated. The predictions generated by the CNN are then utilized as input features for the KNN model.

For the KNN component, various values of k are evaluated to identify the optimal setting for accuracy. Upon determining the best value of k, the KNN model is trained and subsequently evaluated. The performance of the combined CNN and KNN model is assessed through multiple metrics, including the confusion matrix, ROC curve, precision-recall curve, and accuracy as a function of k. The confusion matrix is presented as a heat map, effectively illustrating the correct and incorrect predictions for both the "Apnea" and "Non-Apnea" classes.

The results indicate that the integration of CNN and KNN significantly enhances performance when compared to CNN alone. Initially, The CNN algorithm achieved an accuracy of 89.16%. However, after incorporating the KNN model, the overall accuracy improved to 90.30%. This enhancement underscores the KNN's role in optimizing the results derived from the CNN. By utilizing the CNN's predictions as input features for KNN, we were able to extract more meaningful information from the data, thereby improving classification accuracy.

| | Apnea | No-Apnea | Macro-avg | Weighted-avg |
|---|---|---|---|---|
| Precision | 0.95 | 0.82 | 0.88 | 0.90 |
| Recall | 0.89 | 0.91 | 0.90 | 0.90 |
| F1-Score | 0.92 | 0.86 | 0.89 | 0.90 |

**Table 2:** Performance Evaluation of the Sleep Apnea Detection Model (CNN+KNN).

Table 2 illustrates the performance evolution of the sleep apnea detection model that utilizes a CNN architecture in conjunction with a KNN algorithm. This evaluation focuses on the ability of these models to accurately classify ECG samples into their respective classes: apnea and non-apnea.

- Precision: The combination of the CNN and KNN models achieved a precision of 0.95 for the "Apnea" class, indicating that 95% of the samples classified as apnea are indeed correct. For the "No-Apnea" class, the precision stands at 0.82, reflecting that 82.65% of the samples classified as non-apnea are accurately identified.
- Recall: The recall for the "Apnea" class is 0.89, meaning that 89.20% of actual apnea samples were correctly identified. In contrast, the recall for the "No-Apnea" class is 0.91, indicating that 91.04% of actual non-apnea samples were correctly classified.
- F1-Score: The F1-score, which balances precision and recall, is 0.92 for the "Apnea" class and 0.86 for the "No-Apnea" class.
- Macro-Average: The macro-average metrics provide an overall perspective without considering class imbalance, with a precision of 0.88, a recall of 0.90, and an F1-score of 0.89.
- Weighted Average: The weighted averages, which consider the distribution of classes, yield a precision of 0.90, a recall of 0.90, and an F1-score of 0.90.

These metrics collectively demonstrate the effectiveness of the CNN+KNN model in accurately detecting sleep apnea and highlight the model's strengths in differentiating between apnea and non-apnea samples.

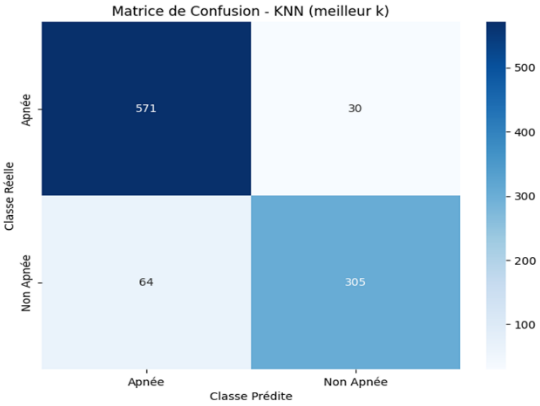


**Figure 4:** The Confusion Matrix of the CNN Model.

Figure 4 demonstrates the significant enhancements achieved when the CNN is integrated with the KNN model, utilizing the optimal k value identified during experimentation. The confusion matrix reveals a notable reduction in both false positives and false negatives, with only 64 false positives and 30 false negatives recorded. This improvement signifies the combined model's enhanced ability to accurately classify data.

Furthermore, the overall precision of the model has increased to 90.30%.

In summary, integrating the KNN model with the CNN has led to significant performance improvements, reducing classification errors and elevating overall accuracy. This suggests that the KNN effectively compensates for certain limitations of the CNN, thereby bolstering the combined model's capability to accurately identify periods of apnea.

## 8. Comparison of Our Results

| Method | Precision | Recall | AUC | F1-Score |
|---|---|---|---|---|
| KNN | 0.77 | 0.77 | 0.84 | 0.80 |
| CNN | 0.89 | 0.89 | 0.88 | 0.85 |
| CNN + KNN | 0.90 | 0.90 | 0.90 | 0.90 |

**Table 3:** Comparison of Our Results.

The results from various models for classifying apnea periods clearly indicate the substantial advantages of the combined CNN-KNN approach. When using the k-Nearest Neighbors (KNN) algorithm alone, precision reaches 77.76%, revealing moderate performance in classifying apnea periods. In contrast, the Convolutional Neural Network (CNN) significantly enhances precision to 89.16%, demonstrating its capability to extract pertinent features from the input data.

Moreover, incorporating KNN to utilize the predictions made by the CNN further boosts precision, achieving 90.30%. This finding highlights KNN's crucial role in optimizing the outcomes generated by the CNN by harnessing the features extracted for enhanced classification. The hybrid approach effectively combines the strengths of both models leveraging the CNN's ability to learn complex representations alongside the simplicity and flexibility of KNN in classification.

## 9. Comparison to Related Works

Comparing our results with those from related works presented in the table below, it is clear that our "CNN-KNN" combined approach occupies a commendable position among previously studied methods regarding the classification accuracy of apnea periods. This synergistic model effectively utilizes the features extracted by the CNN to enhance KNN's classification capabilities. Consequently, our approach presents a promising and competitive solution in the domain of apnea detection within ECG data.

| Study & Method | Precision |
|---|---|
| Study 1: CNN | 97.1% |
| Study 2: CNN | 94% |
| Study 5: 1D CNN | 97.1% |
| Our Study: CNN | 89.16% |
| Study 1: KNN | 82.9% |
| Our Study: KNN | 77.76% |
| Study 3: TW-MLP | 87.3% |
| Study 4: SE-ResNext 50 | 90.28% |
| Our Study: CNN+KNN | 90.30% |

**Table 4:** Comparison with Related Works.

The results presented in Table 4 indicate that our precision rates with CNN and KNN are approximately 89.16% and 77.76%, respectively, which aligns with several prior studies. For instance, Study 1 achieved a precision of 97.1% using CNN, while Study 1 utilizing KNN reached 82.9%. Our findings fall within comparable precision ranges to these earlier works, thereby confirming the robustness of our results.

What sets our approach apart, however, is the implementation of the CNN-KNN combination, which yielded a precision of 90.30%. To the best of our knowledge, this represents the first application of such a combined approach for classifying ECG data in the context of sleep apnea, underscoring the originality and effectiveness of our method. Consequently, our results highlight the value of integrating CNN and KNN models to significantly enhance classification performance in this specialized domain.

## 10. Conclusion

The work represents a contribution to sleep apnea detection, it offers a non-invasive and effective solution that has the potential to enhance the quality of life for patients. The advancements achieved in this work pave the way for new applications of machine learning in healthcare, especially in the diagnosis and management of sleep disorders.

Future prospects are promising, hinting at even more innovative and effective developments, highlighting the ongoing significance of research and innovation in the healthcare field.

### References

[1] Hung-Yu Chang, Cheng-Yu Yeh, Chung-Te Lee, and Chun-Cheng Lin, "A Sleep Apnea Detection System Based on a One-Dimensional Deep Convolution Neural Network Model Using Single-Lead Electrocardiogram," 2020. doi:10.3390, s20154157.

[2] Julie Vicat, "Syndrome d'apnées du sommeil, médicaments du système nerveux central : rôle du pharmacien d'officine," Sciences pharmaceutiques, 2016.

[3] Nguyen A, Nguyen T, Le H, Pham H, and Do C. "A novel deep learning-based approach for sleep apnea detection using single-lead ECG signals," arXiv:2208.03408v2.

[4] Tao Wang, Changhua Lu, Guohao Shen, "Detection of Sleep Apnea from Single-Lead ECG Signal Using a Time Window Artificial Neural Network," 2019.

[5] Tao W, Changhua L, Guohao S, Feng H, "Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network," Academic editor Dennis Lau, 2019.

[6] T Penzel, GB Moody, RG Mark, AL Goldberger, JH Peter. The Apnea-ECG Database. Computers in Cardiology 2000, 27:255-258, IEEE.

[7] Yannick GOTTWALLES, "L'E.C.G. pour les nuls," Directeur Médical de Pôle, Hôpitaux Civils de Colmar, 2008.

## A. Online Resources

[W1] Georges M, Roger M. "Détection et quantification de l'apnée sur la base de l'ECG : The PhysioNet/Computing in Cardiology Challenge 2000." https://physionet.org/content/challenge-2000/1.0.0/

[W2] "Sleep Apnea," https://www.elsan.care/fr/pathologie-et-traitement/maladies-generale/apnee-du-sommeil-causes-traitements

# A Fuzzy-KNN approach for Heart disease Prediction

Amal MILOUD AOUIDATE[1,*,†], Liza AIT HAMOU[1,†]

[1] *Chadli Benjdid University, El-Tarf, Algeria*

**Abstract**

Cardiovascular diseases pose a significant threat to human life, making effective prediction and treatment essential. Artificial intelligence has emerged as a crucial tool in diagnosing these conditions, particularly through its powerful technique: machine learning. This work was done with the aim of creating a system for predicting heart diseases by utilizing the concepts of the K-Nearest Neighbors (KNN) supervised learning algorithm as well as fuzzy logic. This is achieved by creating a symbiosis between these two concepts, allowing for the construction of a two-step model. According to our study, we found that the proposed algorithm provides an excellent trade-off between performance and classification time for predicting heart diseases.

**Keywords**

Machine learning, heart diseases, cardiovascular disease, prediction, KNN, fuzzy logic.

## 1. Introduction

The process of diagnosing and forecasting the risks of heart diseases based on traditional methods—that is, relying on the intuition, knowledge, and experience of the physician—has become insufficient due to the vast amount of data, the high complexity of medical problems, and the limited capabilities of the physician to solve certain medical issues, among other factors. All these elements contribute to diagnostic errors in heart diseases, making it necessary to use new analysis techniques to create models that facilitate decision-making.

Advances in artificial intelligence create new opportunities for improving medical decision-making. This is due to the ability of AI algorithms to analyze large amounts of data that often elude the human eye.

Our main objective in this work is to create a system for predicting heart diseases by introducing the selective features of the K-Nearest Neighbors (KNN) supervised learning algorithm to the strong classification features of fuzzy logic. The combination of these two techniques will improve the prediction of heart diseases from the data, in order to find the best model for an optimal outcome.

## 2. Related works

The literature shows that systems based on fuzzy logic are frequently used in the diagnosis of diseases, where various factors influence the decision-making process and lead to divergences in practitioners' opinions. The application of fuzzy logic and its effectiveness in the medical diagnosis of ankylosing spondylitis, anemia, dengue, thyroid disorders, Alzheimer's disease, blood pressure, diabetes, and mental health has already been demonstrated, as well as in food and medication recommendation systems assisted by an ontology for patients, etc., to identify various disorders [01, 02, 03, 04, 05, 06, 07, 08]. Similarly, numerous research efforts are underway to create a precise, economical, and effective fuzzy logic-based system for the diagnosis of heart diseases, as the World Health Organization (WHO) has reported that cardiovascular diseases (CVD) are now the leading

[1*] Corresponding author.

[†] These authors contributed equally.

a.miloudaouidate@univ-eltarf.dzi (A. Miloud Aouidate);liza.aithamou12@gmail.com (L. Ait Hamou)

cause of death worldwide [01, 09, 10, 11, 12]. Notably, the risk of mortality can be reduced through early diagnosis of heart diseases. However, adaptive algorithms are needed for earlier prediction. Some of these are presented in [13, 14]. A group of researchers in [15, 16, 17, 18, 19, 20, 21, 22, 23, 24] is working on predicting heart diseases. Another group of researchers in [25, 26, 27, 28, 29, 30, 31, 32] focuses on the diagnosis of heart diseases. With a different number of input and output attributes, the authors of [33, 34, 35, 36, 37, 38, 39, 40, 41] present varying levels of accuracy percentages. The accuracy obtained ranges from 63.24% [40] to 94.05% [37] based on various input and output features implemented in Mamdani inference systems. The literature [33] reports an accuracy of 94% using 44 rules in the fuzzy expert system with eleven inputs and one output. The accuracy is slightly increased (94.05%) in [37] as the work introduces the decision tree algorithm with the fuzzy expert system. In recent years, some research works [12, 42, 43, 44, 45] have presented further improvements in accuracy through a hybrid approach of fuzzy algorithms and neural networks, but this hybridization makes the overall system complicated and computationally intensive.

# 3. Proposed model

The proposed model is distinguished by two main steps: first, the refinement of data through the selection of relevant features for the prediction of heart diseases. The second part is the creation of a fuzzy model for the classification and prediction of heart diseases, using the refined dataset obtained from the first step as input.

## 3.1. Features Selection

The k-nearest neighbors (KNN) algorithm is a classification algorithm that relies on the principle of searching for the k closest neighbors. In the context of predicting heart diseases, patient characteristics such as age, sex, cp, and other features are used to determine the nearest neighbors.

We utilized this algorithm to select relevant features that allow for correct data classification. To achieve this, we focused on evaluating the two main parameters of the algorithm, which are distance and the number of neighbors.

Based on the nature of the problem at hand, we chose to use Euclidean distance. This distance measures the direct distance between two points in a straight line.

We opted for this distance because its concept particularly aligns with our problem, where the distance between two points will be the sum of the differences between the values of each feature. This sum should approach zero for the points to be identical. The smaller this distance, the closer the points are, whether graphically or literally. To introduce the principle of features selection, we varied the value of 'n' by changing the features chosen for distance calculation in order to select the most relevant ones.

The model has selected seven features with an overall accuracy of 78%, which is considered acceptable for now. These features are: 'cp', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal'.

To define the appropriate number of neighbors 'K', we proceeded through testing, evaluating the accuracy on the processed heart disease dataset provided by KNN by adjusting 'K' until the accuracy plateaued. After several tests, we concluded that the most appropriate K would be k=7 with an accuracy of 92%.

## 3.2. Prediction

For the prediction of heart diseases, we use the fuzzy logic model, which has allowed us, after testing, to achieve good accuracy. Fuzzy logic consists of three main steps: fuzzification, fuzzy inference, and defuzzification.

### 3.2.1. Fuzzification

Fuzzification is the first step in the fuzzy reasoning process, where actual input values are converted into degrees of membership in fuzzy sets. We applied this method to the set of attributes derived from preprocessing using the KNN algorithm, to assign a degree of membership to each category.

- *Fuzzification of Chest Pain Functions (cp):* There are four fuzzy sets: "typical angina," "atypical angina," "non-anginal pain," and "asymptomatic."
- *Fuzzification of maximum heart rate achieved (thalach):* The fuzzy membership functions for the maximum heart rate achieved consist of three fuzzy sets: "Low," "Normal," and "High."
- *Fuzzification of exercise-induced angina (exang):* The fuzzy membership functions for exercise-induced angina include two fuzzy possibilities: "No" and "Yes."
- *Fuzzification of ST depression induced by exercise compared to rest (oldpeak):* The fuzzy membership functions for ST segment depression induced by exercise consist of two fuzzy sets: "Low" and "Significant."
- *Fuzzification of peak exercise ST segment slope (slope):* The fuzzy membership functions for classifying peak exercise ST segment slope into three categories: Ascending, Flat, and Descending. The fuzzy membership values range from 0 to 1, indicating the degree of membership of a given slope value to each category.
- *Fuzzification of the number of major vessels colored by fluoroscopy (ca):* The fuzzy membership functions for classifying the number of major vessels colored by fluoroscopy into four categories: Zero, One, Two, and Three or more. Again, the membership values represent the degree of membership to each category based on the observed number of vessels.
- *Fuzzification of thalassemia (Thal):* The fuzzy membership functions for classifying thalassemia (a blood disorder) into three categories: Normal, Fixed Defect, and Reversible Defect. The membership values indicate the degree of membership of the patient's condition to each category.

### 3.2.2. Fuzzy Inference

After studying diagnostic models followed by cardiology practitioners, and after establishing the different categories and factors defining the various possibilities of disease, and finally after examining the relationships between the factors that allow us to define a sick patient from a healthy patient, we created a set of fuzzy rules, some examples of which are:

1. If cp is high and thalach is low, then the risk of heart disease is high.
2. If exang is 1 and oldpeak is high, then the risk of heart disease is high.
3. If slope is low and ca is high, then the risk of heart disease is moderate.
4. If thal is normal and cp is low, then the risk of heart disease is low.
5. If thalach is high and exang is 0, then the risk of heart disease is low.
6. If oldpeak is low and slope is medium, then the risk of heart disease is moderate.

### 3.2.3. Defuzzification

The defuzzification method we used for the proposed work is the centroid method (center of gravity), which identifies the balance point of the resulting fuzzy distribution to determine the final risk.

## 4. Results

To evaluate our approaches, it has been essential to rely on well-established metrics, including:

- *Accuracy:* This is the proportion of correct predictions compared to the total number of predictions.
- *Precision:* Measures the number of true positive predictions relative to the total number of positive predictions.
- *Recall (or Sensitivity)*: Assesses the ability to identify all positive instances in a dataset.

- *F1-score:* A balance between precision and recall, useful when you need an equilibrium between the two.

After training the proposed models for heart disease prediction, we obtained the following results: The rule-based fuzzy logic algorithm we previously described achieved an accuracy of 90%. This accuracy is quite acceptable for the detection of cardiovascular diseases, as it encourages the physician to conduct further explorations before dismissing the possibility of disease. When we combined the two concepts, fuzzy logic and K-NN, into a single process, we were able to achieve an accuracy of 95.60%. These results show that the KNN algorithm can serve as an optimizer for a fuzzy logic-based model. KNN improves the model's accuracy .

| Models | Accuracy | Precision | Recall | F1_score |
|---|---|---|---|---|
| Fuzzy Logic | 90.00% | 91.00% | 91.00% | 91.00% |
| Fuzzy / KNN | 95.60% | 96.00% | 96.00% | 96.00% |

**Table 1:** Performance Evaluation Results

Table 1 presents the results of the performance evaluation metrics (accuracy, Precision, Recall, F1-score) for each algorithm: table but remove the KNN row.

In the following (Table 2), we will compare the results we obtained with those presented by researchers whose algorithms have been cited above.

| Test conducted by | Year of publication | Type of method | Accuracy |
|---|---|---|---|
| Literature [09] | 2016 | Diagnosis:Fuzzy Type 2 | 73.78% |
| Literature[10] | 2015 | Prediction:Fuzzy genetic hybrid | 86.00% |
| Literature [24] | 2014 | Diagnosis:Fuzzy model | 88.79% |
| Literature [32] | 2010 | Diagnosis:Fuzzy model | 94.00% |
| Literature [37] | 2016 | Diagnosis:Fuzzy Model | 94.05% |
| Literature [40] | 2019 | Diagnosis:Fuzzy Model | 91.00% |
| Literature [44] | 2020 | Diagnosis:Fuzzy Model | 96.60% |
| Literature [41] | 2017 | Diagnosis:Fuzzy model | 97.78% |

| | | | |
|---|---|---|---|
| Our Fuzzy Model | 2024 | Prediction:Fuzzy model | 90.00% |
| Our KNN/Fuzzy Model | 2024 | Prediction:Hybrid (KNN/ Fuzzy) | 95.60% |

**Table 2:** Comparison to the results obtained by the researchers.

The table 2 shows that our proposed algorithm ranks respectively among the most recent fuzzy logic-based models in heart disease prediction. From these results, we can conclude that our model offers a classification accuracy that surpasses most algorithms seen in the literature and approaches the best algorithm in the literature.

We would like to remind that the purpose of the work is not only diagnosis but especially prediction, which helps to prevent or alleviate the effects of the disease.

## Conclusion

The prediction of heart diseases using machine learning techniques is a promising field in modern medicine. After a long period of work, we have finally completed our thesis on machine learning for predicting heart diseases. Our work involves creating a heart disease prediction system using the supervised learning algorithm K-Nearest Neighbors (KNN) and fuzzy logic, as well as combining fuzzy logic with KNN, and comparing the performance of the algorithms to choose the most effective one. According to our research, fuzzy logic achieved an accuracy of 90% when the KNN improved it, making it able to achieve an accuracy of 95.6% with an acceptable response time.

## References

[1] .M. Maftouni, I. Turksen, M. F. Zarandi, F. Roshani, Type-2 fuzzy rule- based expert system for ankylosing spondylitis diagnosis, in: 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC), IEEE, 2015, pp. 1–5.

[2] .M. F. Shaik, M. M. Subashini, Anemia diagnosis by fuzzy logic using lab- view, in:2017 International Conference on Intelligent Computing and Control (I2C2), IEEE, 2017,pp. 1–5.

[3] .D. Saikia, J. C. Dutta, Early diagnosis of dengue disease using fuzzy inference system, in: 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, 2016, pp. 1–6.

[4] .S. A. Biyouki, I. Turksen, M. F. Zarandi, Fuzzy rule-based expert system for diagnosis of thyroid disease, in: 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2015, pp. 1–7.

[5] .El-Sappagh S., Ali F., Abuhmed T., Singh J., Alonso J. M., Automatic detection of Alzheimer's disease progression: An efficient information fusion approach with heterogeneous ensemble classifiers, Neurocomputing, 2022 Nov 1; 512:203–24.

[6] .El-Sappagh S, Saleh H., Sahal R., Abuhmed T., Islam S. R., Ali F., et al, Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data, Future Generation Computer Systems. 2021 Feb 1; 115:680–99.

[7] .Ali F., El-Sappagh S., Islam S. R., Ali A., Attique M., Imran M., et al, An intelligent healthcare monitoring framework using wearable sensors and social networking data, Future Generation Computer Systems. 2021 Jan 1; 114:23–43.

[8] .Ali F., Islam S. R., Kwak D., Khan P., Ullah N., Yoo S. J., et al, Type-2 fuzzy ontology–aided recommendation systems for IoT–based healthcare. Computer Communications, 2018 Apr 1; 119:138–55.

[9] .Lafta H. A., Oleiwi W. K., A fuzzy petri nets system for heart disease diagnosis, Journal of Babylon University/Pure and Applied Sciences 25 (2) (2017) 317–328.

[10] .Sajiah A. M., Setiawan N. A., Wahyunggoro O., Interval type-2 fuzzy logic system for diagnosis coronary artery disease, Communications in Science and Technology 1 (2).

[11] .Santhanam T., Ephzibah E., Heart disease prediction using hybrid genetic fuzzy model, Indian Journal of Science and Technology 8 (9) (2015) 797.

[12] .Muhammad L., Algehyne E. A., Fuzzy based expert system for diagnosis of coronary artery disease in nigeria, Health and Technology 11 (2) (2021) 319–329. pmid:33614390

[13] .Kaur J., Khehra B. S., Fuzzy logic and hybrid based approaches for the risk of heart disease detection: State-of-the-art review, Journal of The Institution of Engineers (India): Series B (2021) 1–17.

[14] .Reddy G. T., Reddy M. P. K., Lakshmanna K., Rajput D. S., Kaluri R., Srivastava G., Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis, Evolutionary Intelligence 13 (2) (2020) 185–196.

[15] .Gadekallu T. R., Khare N., Cuckoo search optimized reduction and fuzzy logic classifier for heart disease and diabetes prediction, International Journal of Fuzzy System Applications (IJFSA) 6(2) (2017) 25–42.

[16] .Wiharto W., Kusnanto H., Herianto H., Interpretation of clinical data based on c4.5 algorithm for the diagnosis of coronary heart disease, Health- care informatics research 22 (3) (2016) 186–195. pmid:27525160

[17] .M. Tarawneh, O. Embarak, Hybrid approach for heart disease prediction using data mining techniques, in: International Conference on Emerging Internetworking, Data & Web Technologies, Springer, 2019, pp. 447–454.

[18] .Tan C. H., Tan M. S., Chang S. W., Yap K. S., Yap H. J., Wong S. Y., Genetic algorithm fuzzy logic for medical knowledge-based pattern classifi- cation, Journal of Engineering Science and Technology 13 (2018) 242–258.

[19] .Gadekallu T. R., Gao X.-Z., An efficient attribute reduction and fuzzy logic classifier for heart disease and diabetes prediction, Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 14 (1) (2021) 158–165.

[20] .V. S. Dehnavi, M. Shafiee, The risk prediction of heart disease by using neuro-fuzzy and improved goa, in: 2020 11th International Conference on Information and Knowledge Technology (IKT), IEEE, 2020, pp. 127–131.

[21] .L. P. Koyi, T. Borra, G. L. V. Prasad, A research survey on state of the art heart disease prediction systems, in: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, 2021, pp. 799–806.

[22] .Liu Y., Eckert C. M., Earl C., A review of fuzzy ahp methods for decision- making with subjective judgements, Expert Systems with Applications (2020) 113738.

[23] .Wadhawan S., Maini R., A systematic review on prediction techniques for cardiac disease, International Journal of Information Technologies and Systems Approach (IJITSA) 15 (1) (2022) 1–33.

[24] .Hameed A. Z., Ramasamy B., Shahzad M. A., Bakhsh A. A. S., Efficient hybrid algorithm based on genetic with weighted fuzzy rule for developing a decision support system in prediction of heart diseases, The Journal of Supercomputing (2021) 1–21

[25] .Devi Y. N., Anto S., An evolutionary-fuzzy expert system for the diagnosis of coronary artery disease, Int. J. Adv. Res. Comput. Eng. Technol 3 (4) (2014) 1478–1484.

[26] .Wiharto W., Kusnanto H., Herianto H., Hybrid system of tiered multi-variate analysis and artificial neural network for coronary heart disease diagnosis, International Journal of Electrical and Computer Engineering 7 (2) (2017) 1023.

[27] .Priyatharshini R., Chitrakala S., A self-learning fuzzy rule-based system for risk-level assessment of coronary heart disease, IETE Journal of Research 65 (3) (2019) 288–297.

[28] .Song J., Ni Z., Jin F., Li P., Wu W., A new group decision-making approach based on incomplete probabilistic dual hesitant fuzzy preference relations, Complex & Intelligent Systems 7 (6) (2021) 3033–3049.

[29] .Akhoondi R., Hosseini R., Mazinani M., A GA approach for tuning membership functions of a fuzzy expert system for heart disease prognosis development risk, Journal of Computing and Security 4 (1) (2017) 13–23.

[30] .Eisa M. M., Alnaggar M. H., Hybrid rough-genetic classification model for IoT heart disease monitoring system, in: Digital Transformation Technology, Springer, 2022, pp. 437–451.

[31] .Nazari S., Fallah M., Kazemipoor H., Salehipour A., A fuzzy inference- fuzzy analytic hierarchy process-based clinical decision support system for diagnosis of heart diseases, Expert Systems with Applications 95 (2018) 261–271.

[32] .Hern, Framework for the development of data-driven Mamdani-type fuzzy clinical decision support systems.

[33] .Adeli A., Neshat M., A fuzzy expert system for heart disease diagnosis, in Proceedings of the international multi conference of engineers and computer scientists, Hong Kong, Vol. 1, Citeseer, 2010, pp. 28–30.

[34] .Khatibi V., Montazer G. A., A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment, Expert Systems with Applications 37 (12) (2010) 8536–8542.

[35] .Kim J.-K., Lee J.-S., Park D.-K., Lim Y.-S., Lee Y.-H., Jung E.-Y., Adaptive mining prediction model for content recommendation to coronary heart disease patients, Cluster computing 17 (3) (2014) 881–891.

[36] .W. M. Baihaqi, N. A. Setiawan, I. Ardiyanto, Rule extraction for fuzzy expert system to diagnose coronary artery disease, in: 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, 2016, pp. 136–141.

[37] .V. Krishnaiah, G. Narsimha, N. S. Chandra, Heart disease prediction system using data mining technique by fuzzy k-nn approach, in: Emerging ICT for Bridging the FutureProceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1, Springer, 2015, pp. 371–384.

[38] .Mokeddem S., Atmani B., Assessment of clinical decision support systems for predicting coronary heart disease, International Journal of Operations Research and Information Systems (IJORIS) 7(3) (2016) 57–73.

[39] .T. Kasbe, R. S. Pippal, Design of heart disease diagnosis system using fuzzy logic, in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE, 2017, pp. 3183–3187.

[40] .O. Terrada, B. Cherradi, A. Raihani, O. Bouattane, A fuzzy medical diagnostic support system for cardiovascular diseases diagnosis using risk factors, in: 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), IEEE, 2018, pp. 1–6.

[41] .P. Jain, A. Kaur, A fuzzy expert system for coronary artery disease diagnosis, in Proceedings of the Third International Conference on Advanced Informatics for Computing Research, 2019, pp. 1–6.

[42] .Uyar K., Ilhan A., Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks, Procedia computer science 120 (2017) 588–593.

[43] .Iancu I., Heart disease diagnosis based on mediative fuzzy logic, Artificial intelligence in medicine 89 (2018) 51–60. pmid:29859751

[44] .H. Kahtan, K. Z. Zamli, W. N. A. W. A. Fatthi, A. Abdullah, M. Ab- dulleteef, N. S. Kamarulzaman, Heart disease diagnosis system using fuzzy logic, in: proceedings of the 2018 7th International Conference on Software and Computer Applications, 2018, pp. 297–301.

[45] .P. Krishnan, V. Rajagopalan, B. I. Morshed, Anovel severity index of heart disease from beat-wise analysis of ECG using fuzzy logic for smart-health, in: 2020 IEEE International Conference on Consumer Electronics (ICCE), IEEE, 2020, pp. 1–5.

# A Machine Learning Algorithms for decision-makers and analysts in the field of digital finance*

Atmane Hadji[1,*],Farid Boumaza[2,3]

[1]LISI Laboratory, Computer Science Department, University Center A. Boussouf Mila, 43000 Mila, Algeria
[2] Computer Science Department, University of Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj 34030, Algeria
[3]LAPECI Laboratory , University of Oran1, Oran 31000, Algeria.

**Abstract**
Sentiment analysis on social networks has emerged as a critical field of study in recent years, driven by the widespread adoption of these platforms and the growing volume of user-generated data. Understanding online opinions and sentiments is essential for deriving valuable insights into public attitudes and identifying social and economic trends. This study focuses on leveraging machine learning techniques for sentiment analysis due to their efficiency in handling large and diverse datasets. Four popular algorithms were employed: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest. Among these, SVM demonstrated superior accuracy, making it the optimal choice for sentiment detection in this context. The proposed approach was applied to public discussions about Bitcoin on Facebook, a topic of significant interest in financial and economic domains. The results demonstrated that machine learning techniques could extract precise sentiments and provide a comprehensive understanding of economic trends, offering valuable insights for decision-makers and analysts in the field of digital finance. The findings underline the effectiveness of machine learning methods in sentiment analysis and highlight their potential in processing diverse types of data from social media platforms. These methods can be extended to other domains, such as e-commerce, consumer opinion analysis, and market research, reinforcing their value as tools for strategic decision-making across various sectors. Furthermore, this study illustrates the adaptability of machine learning techniques to handle the unique characteristics of large and heterogeneous datasets, paving the way for the development of more advanced analysis systems. These systems could address other areas, including news sentiment analysis, global events, and even cultural and social content, showcasing the vast opportunities offered by these technologies.

**Keywords**
Social Networks, Bitcoin, Machine learning ,SVM,K-NN,Decision Tree,Random Forest, CEUR-WS

## 1. Introduction

In recent years, social media platforms have become indispensable spaces where individuals express their opinions, sentiments, and experiences, generating an enormous volume of textual data ripe for analysis. This explosion of user-generated content has fueled the advancement of sentiment analysis, a rapidly evolving field focused on extracting, interpreting, and categorizing emotions and opinions embedded in textual data. Sentiment analysis finds extensive applications across diverse domains, including marketing, finance, economics, and politics, where it facilitates understanding public attitudes, identifying trends, and aiding strategic decision-making. For example, in the economic context, sentiment analysis is instrumental in

* The 2nd International Conference on Autonomous Systems and their Applications (ICASA'24) , December15–15, 2024, Et Tarf, Algeria.
*Corresponding author.

a.hadji@centre-univ-mila.dz (A. Hadji); farid.pgia@gmail.com (F. Boumaza)
0000-0001-6706-6360 (A. Hadji); https://orcid.org/0000-0002-9785-420X (F. Boumaza)

decoding consumer and investor perceptions, enabling organizations to anticipate market behaviors effectively.

Despite its promise, extracting opinions accurately from vast, unstructured textual datasets presents significant challenges. Traditional rule-based and static indexing methods often struggle to capture the complexity and contextual nuances inherent in natural language. The variability in expression, sarcasm, implicit sentiment, and evolving terminologies further complicate this task. To overcome these challenges, machine learning-based approaches have emerged as powerful solutions capable of adapting to diverse datasets and recognizing sentiment patterns with high precision.

Unlike static models, machine learning leverages algorithms that learn from data to identify patterns and adapt dynamically. In this study, we utilize state-of-the-art machine learning techniques, including Support Vector Machines (SVM), Decision Trees, Random Forest, and K-Nearest Neighbors (K-NN), to perform sentiment classification. These methods excel in processing large datasets and are particularly effective at handling the variability and ambiguity often found in user-generated content. Among these, SVM has shown superior performance in capturing the subtle interplay between linguistic features and sentiment classes.

This research specifically applies machine learning-based sentiment analysis to discussions surrounding Bitcoin, a topic of significant relevance in the financial domain. The aim is to assess how well these algorithms can extract actionable insights from user opinions and economic discussions. By focusing on this application, we seek to highlight the ability of machine learning approaches to provide nuanced interpretations of public sentiment, enabling a deeper understanding of economic trends and public perceptions.

Ultimately, this study underscores the versatility of machine learning in addressing sentiment analysis challenges and paves the way for its broader application in fields such as consumer opinion analysis, digital marketing, and financial forecasting. By leveraging machine learning's adaptive capabilities, this research contributes to the development of more robust and scalable sentiment analysis systems designed to navigate the complexities of real-world textual data.

## 2. Background and Related works

### 2.1. Machine Background

### 2.1.1. Modifying ML Opinion Extraction

Machine learning-based opinion extraction utilizes trained models on extensive datasets to automatically identify sentiments and opinions in diverse and dynamic contexts. This approach stands out for its adaptability, as it captures the complexities of human language without relying on predefined static rules.

### 2.1.2. Automated Sentiment Classification

Supervised learning models, such as decision trees, support vector machines (SVM), K-Nearest Neighbors (K-NN), and random forest, play a central role in this approach. These models categorize sentiments into predefined classes—positive, negative, or neutral—by learning from labeled training data. Their ability to generalize makes them particularly effective in handling varied linguistic structures and expressions.

### 2.1.3. Pattern Recognition in Text

Unlike traditional rule-based systems, which rely on static and rigid patterns, machine learning models excel in identifying intricate and dynamic patterns within text. By training on diverse datasets, these models learn to detect subtle relationships and nuances in expressed sentiments, including sarcasm, idiomatic phrases, and complex grammatical structures. This allows for a deeper and more accurate understanding of user-generated content.

### 2.1.4. Adaptability and Scalability

A notable advantage of machine learning methods is their adaptability. Models can be retrained with new datasets, enabling them to remain relevant as linguistic trends, topics, or social contexts evolve. This scalability ensures that sentiment extraction remains robust and accurate across varying domains and timeframes.

### 2.1.5. Comparative Evaluation

This study investigates the machine learning-based approach for opinion extraction, focusing on four algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest. The evaluation assesses the effectiveness of these algorithms in extracting opinions from text, particularly in economic and social contexts, such as discussions on social media.

Each of the four machine learning algorithms offers unique advantages in terms of accuracy, pattern recognition, and adaptability. While some may excel at capturing certain types of sentiment or context, others may be more suited to different types of data. By comparing the performance of these algorithms, the study aims to highlight their strengths and limitations in sentiment analysis, providing insights into their practical applications for understanding public opinion and sentiment trends in areas like economics, finance, and social analysis.

## 2.2. Related works

This section explores the state of the art in machine learning-based information extraction (IE) methods, highlighting their flexibility, scalability, and the ability to adapt to dynamic data environments. Machine learning (ML) techniques have gained significant prominence in the field of information extraction due to their capacity to process and analyze unstructured data from diverse sources such as text, images, and documents. Unlike traditional rule-based systems, ML models identify patterns within large datasets, learning from the data to automatically recognize key entities, relationships, and sentiments without predefined rules. This characteristic makes ML-based IE methods particularly valuable for handling complex and ever-evolving datasets across various domains, from healthcare to finance, and social media analysis.

Machine learning methods for information extraction are diverse and include supervised learning models such as Support Vector Machines (SVM), Random Forest, Decision Trees, and deep learning models. These algorithms can be trained on large labeled datasets, enabling them to learn how to classify or extract relevant information based on the patterns found in the data. The flexibility of these algorithms lies in their ability to improve over time as more data becomes available, making them well-suited for applications where data evolves or where new types of information need to be extracted.

ML-based IE methods have been successfully applied across various fields, demonstrating their power in extracting structured data from unstructured sources.

#### 2.2.1. Clinical Data Extraction

Several studies have demonstrated the power of ML techniques in the healthcare domain. A study on clinical data [1] utilized ML and natural language processing (NLP) to identify fracture types in radiology reports, showcasing how ML can convert unstructured medical data into actionable knowledge. In another clinical study [2], ML algorithms were used to analyze radiology reports for detecting abnormalities, highlighting ML's ability to capture contextual information with high accuracy.

#### 2.2.2. Invoice Processing

ML models have been applied to automate invoice processing by using deep learning models such as LayoutLM, which excels in understanding the layout and structure of documents. This study [3] showed that ML can outperform traditional methods in handling layout variations in invoices, thus improving the efficiency of document processing systems.

#### 2.2.3. Misinformation Detection

In the domain of social media and misinformation detection, an ML-based approach was employed to identify "fake news" related to COVID-19. By leveraging medical features, this model [4] enhanced the detection accuracy of misleading information, demonstrating the potential of ML for content verification and sentiment analysis in public health domains.

#### 2.2.4. Document Understanding and Extraction

Recent studies [5],[6] have demonstrated the effectiveness of transformer models, such as BERT, in extracting information from handwritten documents and resumes. These models excel at handling complex unstructured data by capturing semantic and syntactic patterns, making them invaluable tools for extracting key information from diverse document types. Machine learning-based information extraction methods offer several significant advantages over traditional approaches:

#### 2.2.5. Scalability and Adaptability:

ML algorithms can handle vast amounts of data from diverse sources and can be retrained to adapt to new types of data. This makes them ideal for dynamic environments where the nature of the data is constantly changing, such as social media or evolving market trends.

#### 2.2.6. Contextual Understanding:

Unlike rule-based systems that rely on predefined patterns, ML models, especially deep learning models, have the ability to understand context. This is particularly valuable in tasks like sentiment analysis and named entity recognition, where meaning can change depending on the context in which a term or phrase is used.

#### 2.2.7. Precision and Automation:

ML models are capable of automatically identifying and extracting relevant information with high precision, reducing the need for manual intervention. As these models are trained on large datasets, they improve over time, providing more accurate extractions as more data is processed.

## 3. Proposed approach

The following architecture (Figure 1) depicts the detailed design of our opinion analysis system. The proposed system consists of several stages:

### 3.1. Data Collection

As part of this research, a dataset of 1,000 Facebook comments was compiled to analyze sentiments related to Bitcoin. These comments were collected over a six-month period from Facebook pages dedicated to the topic, ensuring adequate temporal and thematic coverage to evaluate the robustness of the results. We obtained the data from online social networks, particularly Facebook, and processed comments related to fan opinions semi-automatically. To efficiently extract relevant comments from popular social media platforms like Facebook and Twitter, we leveraged the GATE platform (General Architecture for Text Engineering). The discussions include various themes such as Bitcoin's price, volatility, and adoption, providing a diverse perspective on user perceptions. The data labeling was performed manually by two independent annotators, an approach aimed at ensuring the quality and reliability of the labels. To minimize biases in the results, a balance was maintained between the sentiment categories (positive, negative, and neutral), with each class representing an almost equal percentage of the dataset. This rigorous methodology ensures the representativeness and validity of the data used in the analysis..

### 3.2. Pretreatment

In this step, we identified the comments related to the Champions League, then processed them in the next step. The filtering techniques applied to the corpus include more than one baseband. We filter the data by bypassing extra spaces and formatting elements to obtain plain text. Consequently, typos are corrected using automated and manual tools, and text normalization is followed, including the removal of special characters, spaces and punctuation [7].

Currently, social media worldwide is considered the most visited source for information on modern technologies like Bitcoin. Bitcoin is the most prominent cryptocurrency with the largest market capitalization. Additionally, it is a digital currency that users can only access online. Thus, online platforms play a crucial role in disseminating information to individuals about Bitcoin and how it is used. People mainly turn to social media when making purchase decisions, including buying or investing in Bitcoin, which is why we chose social media—specifically Facebook, as it gathers all segments of society.

In our study, we reviewed research articles in the field to extract information on factors affecting Bitcoin, which we categorize into three types: positive factors, negative factors, and neutral factors [8].

#### 3.2.1. Positive Factors

We identified several positive factors impacting Bitcoin's increase in value, including but not limited to rising demand, institutional adoption, inflation and economic instability, heightened media coverage, and other elements.

### 3.2.2. Negative Factors

The depreciation of Bitcoin is influenced by multiple factors, some of which include high volatility, economic crises such as wars, high-interest rates, competition from other crypt ocurrencies, difficulty in using it as currency, and additional factors.

### 3.2.3. Neutral Factors

There are also neutral elements, some of which are mentioned below: competition assessment, stability, and media updates.

The goal of extracting these factors that influence Bitcoin's value is to better understand the market and predict future trends, to enhance individuals' confidence in Bitcoin, encourage its usage, expand its application across different fields, improve the performance of exchanges and other platforms, and help more people understand this currency. Additionally, it aims to provide insight into the risks associated with investing in Bitcoin, protecting consumers from fraud.

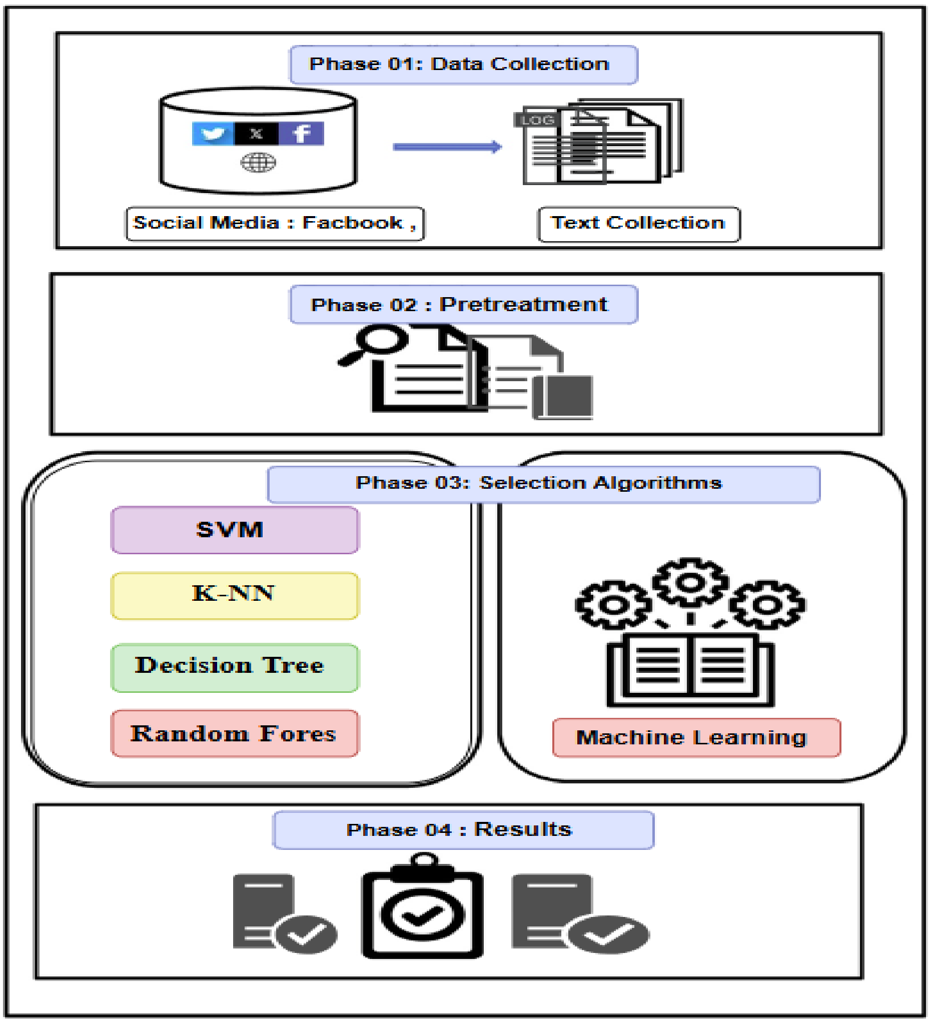


**Figure 1:** General architecture of the proposed system

We also focus on analyzing opinions about Bitcoin through posts and comments on Facebook regarding Bitcoin's price, satisfaction levels, and associated risks. Through this feedback, it is possible to:

- Determine the extent of Bitcoin's popularity;
- Assess whether people are optimistic or pessimistic about its future and better understand their needs;
- Measure public confidence in Bitcoin, their satisfaction level, and future expectations;
- Enable developers to design new technologies to improve market efficiency;
- Facilitate transactions and raise awareness of the risks associated with investing in Bitcoin, as well as provide insight into its influence on the economy and society.

## 4. Selection Algorithms

Machine learning is a field of artificial intelligence that enables computer systems to learn and improve automatically from experience. By using algorithms and mathematical models, it analyzes data to recognize patterns and make decisions without being explicitly programmed.

Machine learning applications are diverse, ranging from speech recognition and online product recommendations to fraud detection and autonomous driving. This field is rapidly advancing due to technological progress and the increasing availability of massive datasets, opening new possibilities across many industrial and scientific sectors [9].

In this study, we explore the application of machine learning for disaster information extraction using four distinct algorithms: Support Vector Machines (SVM), K-Nearest Neighbors, Random Forest Classifier, and Decision Tree Classifier. Each algorithm presents unique characteristics and specific advantages that influence their effectiveness in extracting relevant information:

### 4.1. Support Vector Machines (SVM)

The SVM algorithm is known for its ability to classify data by finding the optimal hyperplane that separates classes with the maximum margin. In the context of natural disasters, it is used to distinguish different categories of information from complex textual data [10].

For linearly separable data, the separation hyperplane can be determined by:

$$WT\,x + b = 0 \qquad (1)$$

- W is the weight vector (or normal) of the hyperplane.
- x is the feature vector of a data point.
- b is the bias (offset) of the hyperplane.

### 4.2. Random Forest Classifier

The Random Forest algorithm improves classification performance by leveraging an ensemble of decision trees. By combining the outputs of multiple trees, it enhances generalization and reduces the risk of overfitting, making it particularly effective for managing diverse and noisy text data [11].

A Random Forest Classifier is an ensemble learning technique that merges the predictions of several decision trees to boost classification accuracy and mitigate overfitting. Each tree is trained on randomly selected subsets of data and features.

$$\hat{y} = \text{mode}\left(\{T_i(\mathbf{x}) \mid i = 1, 2, \ldots, N\}\right)$$

where:

- *p(y=1/x)* = *T_i(\mathbf{x})* is the prediction of the iii-th decision tree,
- **N** is the number of trees,
- The mode function returns the most common class label among all trees' predictions

### 4.3. K-Nearest Neighbors (K-NN)

The K-Nearest Neighbors (K-NN) algorithm is a simple and effective method for classification and regression. It assigns a class to a data point based on the majority class of its kkk-nearest neighbors in the feature space. K-NN is particularly useful for handling multi-class problems and works well when the data distribution is localized.

K-NN classifies a data point by calculating the distances between it and all other points in the dataset, selecting the kkk-closest neighbors, and assigning the majority class label among those neighbors[12].

For a given data point x, the distance to each neighbor is computed using a metric like Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2}$$

Where:

- $x$ is the input feature vector.
- $x_i$ is the feature vector of the i-th neighbor.
- n is the number of features.
- The class of x is determined by the majority vote among the k-nearest neighbors.

### 4.4. Decision Tree Classifier

Decision trees are highly interpretable models that operate by making a series of binary decisions. They are well-suited for extracting straightforward rules from textual data and provide clarity in understanding the criteria used for classification.

Definition: A Decision Tree Classifier divides data into subsets based on specific feature values, constructing a tree-like structure where each node corresponds to a decision guided by an attribute[13].

$$Gini(D) = 1 - \sum_{i=1}^{k} (p_i)^2$$

Where : D is a dataset,

- k is the number of classes,
- pi is the proportion of instances belonging to class i.

The tree continues to split until it reaches a stopping criterion, such as a maximum depth or minimum number of samples per leaf.

## 5. Results and Evaluation

After running the various algorithms with the use of dataset, the system is now able to detect the entities named "Positive Opinion", " Negative Opinion" and "Neutral Opinion" corresponding to opinions on a Cryptocurrency ''Bitcoin''.

The data used in the dataset for the first ontology-based method is the same as that used in the machine learning approach. This dataset is annotated with a range of attributes to support effective information extraction and sentiment analysis, including the classification of sentiments into Positive Opinions, Neutral Opinions, and Negative Opinions.

These annotations aim to evaluate machine learning models designed to extract relevant opinions related to Bitcoin. Figures 3 and 4 illustrate the results obtained for each algorithm used in our study: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest.

To evaluate and compare the methods we studied, we will use metrics: Precision, Recall, and F-scale. Precision refers to the correctness of the retrieval, while recall refers to the completeness of the retrieval. The F-measure provides the harmonic mean between precision and recall [14].
According to [15] :

- Precision is the percentage of correct results among the results obtained .

$$\text{Precision} = \frac{\text{Number of NE correctly recognized}}{\text{Number of NE recognized}}$$

- Recall is the percentage of correct results among the results that must be found .

We present formulas for evaluation such as precision, recall which are widely used measures in NLP evaluations

$$\text{Recall} = \frac{\text{Number of correctly recognized NE}}{\text{Number of corpus NE}}$$

The F-measure is the combination of precision and recall and their weighting. The formula for the F-measure is as follows:

$$\text{F} - \text{mesure} = \frac{2(\text{precision} * \text{ recall})}{\text{precision} + \text{ recall}}$$

The performance of four machine learning algorithms Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Random Forest Classifier (RFC), and Decision Tree Classifier (DTC) was evaluated based on key metrics such as precision, recall, and F-measure. The results for each algorithm are presented in Tables 1, 2, 3, and 4, showcasing the comparative performance of these models in sentiment analysis of Bitcoin-related discussions on Facebook. These tables highlight the strengths and weaknesses of each algorithm, providing insights into their effectiveness for sentiment detection in social media contexts.

**Table 1**

Results of Opinion Extraction in SVM Algorithm

| SVM | Precision | Recall | F-mesure |
|---|---|---|---|
| Negative Opinion | 0,83 | 0 ,91 | 0,87 |
| Neutral Opinion | 0,91 | 1,00 | 0;95 |
| Positive Opinion | 1,00 | 0,57 | 0,73 |
| Weited Average | 0,90 | 0,86 | 0,86 |

**Table 2**

Results of Opinion Extraction in K-NN

| K-NN | Precision | Recall | F-mesure |
|---|---|---|---|
| Negative Opinion | 0,83 | 0 ,91 | 0,87 |
| Neutral Opinion | 0,88 | 0,70 | 0,78 |
| Positive Opinion | 0,57 | 0,57 | 0,57 |
| Weited Average | 0,78 | 0,75 | 0,76 |

**Table 3**

Results of Opinion Extraction in Random Forest Algorithm

| Random Forest | Precision | Recall | F-mesure |
|---|---|---|---|
| Negative Opinion | 0,77 | 0 ,91 | 0,83 |
| Neutral Opinion | 0,91 | 1,00 | 0,95 |
| Positive Opinion | 1,00 | 0,57 | 0,73 |
| Weited Average | 0,88 | 0,86 | 0,85 |

**Table 4**

Results of Opinion Extraction in Decision Tree Algorithm

| Decision Tree | Precision | Recall | F-mesure |
|---|---|---|---|
| Negative Opinion | 0,83 | 0 ,91 | 0,87 |
| Neutral Opinion | 0,83 | 1,00 | 0,91 |
| Positive Opinion | 1,00 | 0,57 | 0,73 |
| Weited Average | 0,88 | 0,86 | 0,85 |

The following table (Table 5) presents the averages of the three metrics (Precision, Recall, F-measure) for the four algorithms used in the study: SVM, K-NN, RFC, and DTC. These averages reflect the overall performance of these algorithms when applied to the dataset used in the analysis

**Table 5**

Average Results of Opinion Extraction in Four Algorithms

| | Precision | Recall | F-mesure |
|---|---|---|---|
| SVM | 0,90 | 0,86 | 0,86 |
| K-NN | 0,78 | 0,75 | 0,76 |
| RFC | 0,88 | 0,86 | 0,85 |
| DTC | 0,88 | 0,86 | 0,85 |
| Average | 0,86 | 0,8325 | 0,83 |

## 6. Analysis and Discussion

### 6.1. Results Analysis

The results obtained in this study are highly satisfactory, as demonstrated by the Precision, Recall and F-mesure (see to Table 1, Table 2 ,Table 3 and Table 4 ).

This section provides an in-depth analysis of the performance of the four algorithms (SVM, K-NN, Random Forest, and Decision Tree) used for opinion detection and sentiment analysis related to Bitcoin, based on data extracted from Facebook. The performance is compared in terms of precision, recall, and F-measure for three categories of opinions: negative, neutral, and positive.

#### 6.1.1. SVM

The SVM algorithm (Figure 2) demonstrates the best overall performance with an average precision of 0.90, recall of 0.86, and F-measure of 0.86.

- **Negative Opinions:** An F-measure of 0.87 reflects a strong ability to detect critical opinions.
- **Neutral Opinions:** Exceptional performance (F-measure of 0.95), with perfect recall (1.00).
- **Positive Opinions:** Although precision is perfect (1.00), the limited recall of 0.57 lowers the overall effectiveness in this category (F-measure of 0.73).

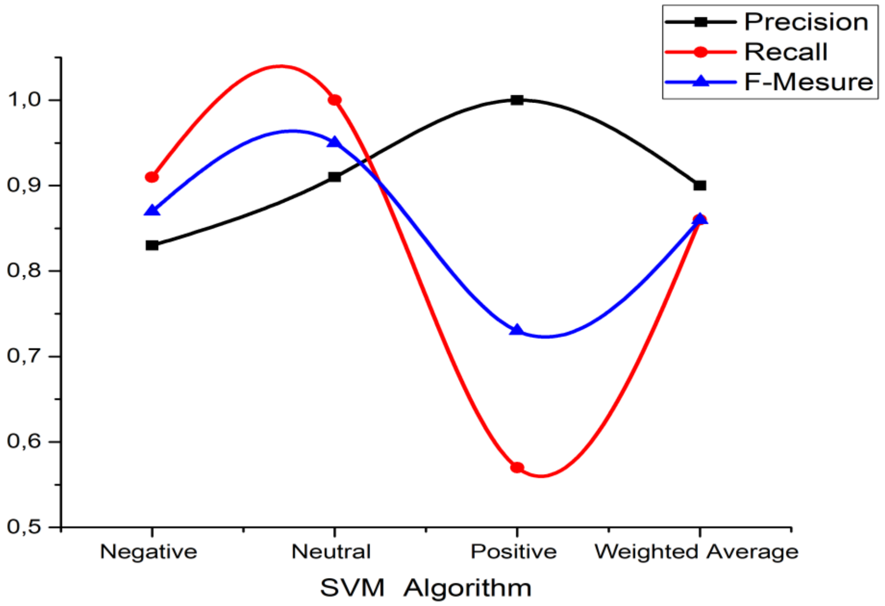


**Figure 2:** Performance Metrics of the SVM Algorithm

### 6.1.2. K-Nearest Neighbors (K-NN)

The K-NN algorithm (Figure 3) is the least effective algorithm, with an average precision of 0.78, recall of 0.75, and F-measure of 0.76.

- **Negative Opinions:** An F-measure of 0.87, comparable to SVM, indicates reasonable capability in detecting these sentiments.
- **Neutral Opinions:** Lower performance (F-measure of 0.78), due to limited recall (0.70).
- **Positive Opinions:** Particularly poor results with an F-measure of 0.57, reflecting challenges in capturing this category accurately.

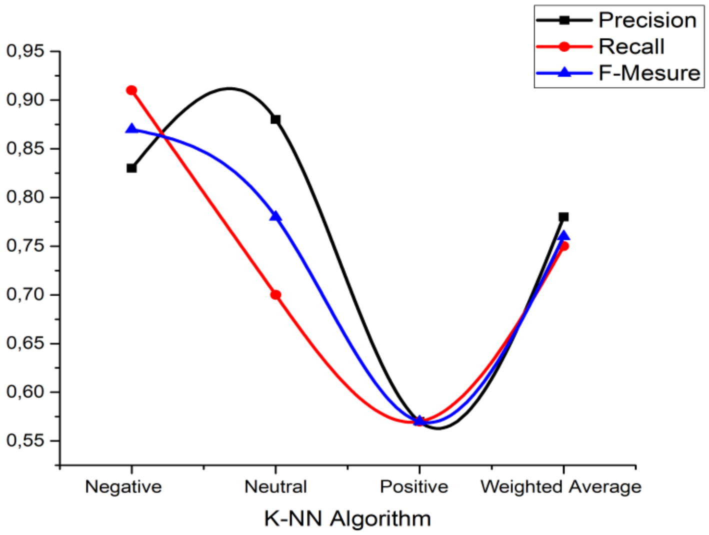


**Figure 3:** Evaluation Metrics for the K-NN Algorithm

### 6.1.3. Random Forest

Random Forest Classifier (Figure 4) achieves strong results overall, with a precision of 0.88, recall of 0.86, and F-measure of 0.85.

- **Negative Opinions:** An F-measure of 0.83, slightly lower than SVM and K-NN.
- **Neutral Opinions:** Excellent performance (F-measure of 0.95), comparable to SVM.
- **Positive Opinions:** Similar to SVM, with perfect precision (1.00) but low recall (0.57), resulting in an F-measure of 0.73.

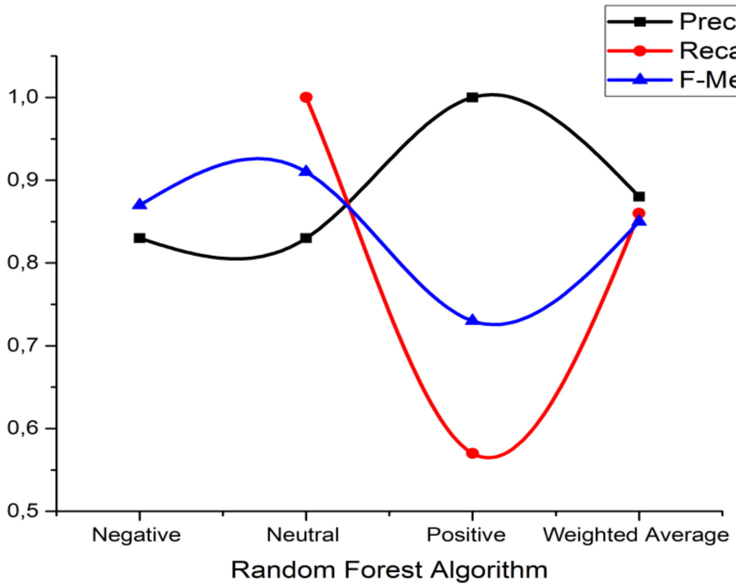


**Figure 4:** Performance Analysis of the Random Forest Algorithm

### 6.1.4. Decision Tree

Decision Tree Classifier (Figure 5) displays comparable performance to Random Forest, with an average precision of 0.88, recall of 0.86, and F-measure of 0.85.

- **Negative Opinions:** An F-measure of 0.87, similar to SVM.
- **Neutral Opinions:** Solid performance (F-measure of 0.91), slightly below that of Random Forest and SVM.
- **Positive Opinions:** As with other algorithms, perfect precision (1.00) but low recall (0.57) limits the F-measure to 0.73.

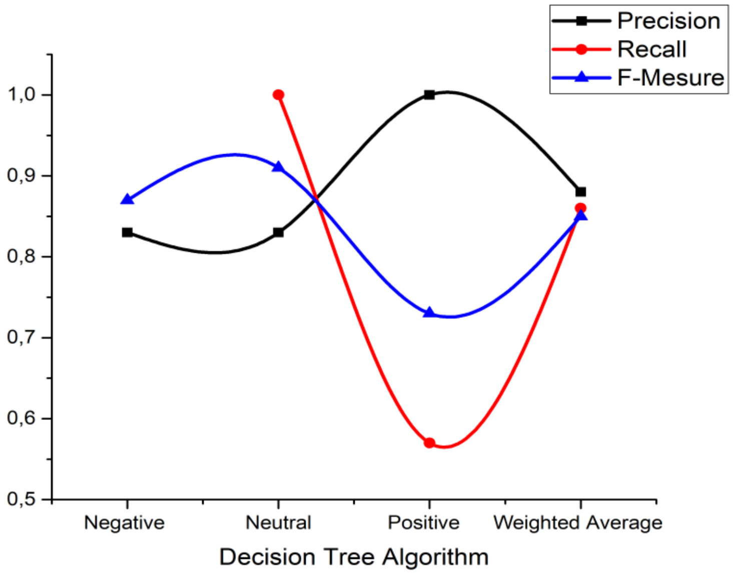


**Figure 5:** Evaluation Metrics for the Decision Tree Algorithm

## 6.2. Comparative Algorithms

- **Overall Performance:**

The following figure (See Figure 6) presents a comparative analysis of the performance metrics (precision, recall, and F-measure) across the four algorithms: SVM, K-NN, RFC, and DTC. The figure visually illustrates the differences in performance between these models, highlighting SVM as the top performer in all evaluated metrics. While RFC and DTC show similar results, they fall slightly behind SVM, particularly in recall and F-measure. K-NN demonstrates the lowest performance, particularly in recall, which emphasizes the model's limitations in handling sentiment analysis for the given dataset. This visual comparison provides a clear understanding of each algorithm's relative strengths and weaknesses.

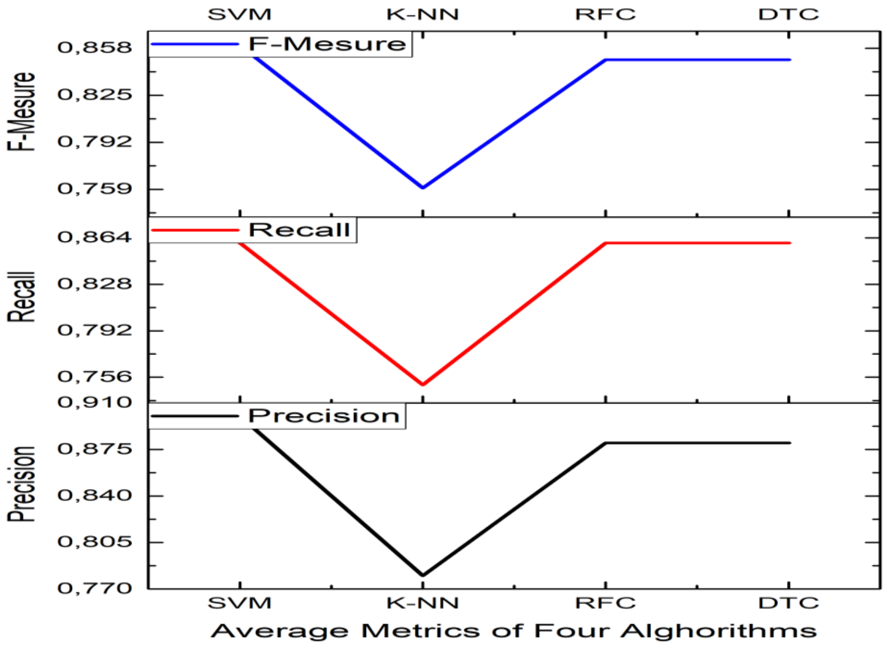


**Figure 6:** Global Performance of the four Algorithms

- SVM is the top-performing algorithm due to its ability to handle complex data and maximize class separation, particularly for neutral and negative opinions.
- K-NN, while intuitive, shows the lowest overall performance, especially for positive opinions, likely due to its sensitivity to noise and inability to capture complex boundaries.
- Random Forest and Decision Tree exhibit similar performances, demonstrating effectiveness in capturing complex patterns through their decision-tree-based approaches.
- Neutral Opinions: All algorithms, except K-NN, perform well in this category. SVM and Random Forest are particularly effective, achieving perfect recall (1.00).
- Positive Opinions: Detecting positive opinions remains a major challenge for all algorithms, with low recall values (0.57). This may be due to data imbalance or difficulty distinguishing positive sentiments in ambiguous texts. The recall shows the weakest overall performance, especially for positive opinions, likely due to its sensitivity to noise and its inability to capture complex boundaries.
- Robustness and Generalization: Tree-based algorithms (Random Forest and Decision Tree) offer greater robustness due to their ability to avoid overfitting, though they fall slightly behind SVM in overall performance.

## 7. Conclusion

This study provides a comprehensive evaluation of machine learning techniques for sentiment analysis of Bitcoin-related discussions on Facebook, with a focus on four algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest. The results demonstrated that SVM outperformed the other algorithms, achieving the highest precision, recall, and F-measure. This finding underscores the effectiveness of machine learning models, particularly SVM, in accurately capturing sentiment nuances in social media contexts.

The implications of these results are significant for financial decision-makers. For instance, identifying user sentiment can aid in predicting real-time Bitcoin price fluctuations, enabling better market analysis. Additionally, sentiment analysis can inform investment strategies, as understanding public opinions can guide decisions on Bitcoin investments. However, ethical considerations must be taken into account to mitigate the risk of biases or inaccuracies in sentiment analysis, which could lead to misguided financial decisions. Implementing mechanisms to detect and address these biases will be essential to ensure the fairness and reliability of these analytical tools.

Looking ahead, several directions for future research can be pursued to further enhance this study's contributions. Expanding the analysis to other cryptocurrencies, such as Ethereum or Ripple, would offer a broader view of the crypto market and its sentiments. Additionally, integrating multilingual data from social media platforms can help capture a wider range of opinions, improving the robustness of the analysis across different regions and cultures. The use of advanced deep learning models like BERT or LSTM could also refine sentiment analysis by better handling complex linguistic subtleties in social media discussions.

These advancements will not only improve the accuracy of sentiment analysis but also broaden its applications to other domains, such as finance, healthcare, and digital commerce. By incorporating these suggestions, future research can significantly enhance the practical impact and relevance of sentiment analysis, paving the way for more informed and reliable decision-making in the digital age.

## References

[1] Fiebeck, J., Laser, H., Winther, H. B., Gerbel, S. : Leaving no stone unturned: using machine learningbasedapproaches for information extraction from full texts of a research data warehouse. In International Conference on Data Integration in the Life Sciences , Cham: Springer International Publishing, pp. 50-58,2018.

[2] Steinkamp, J.M., Chambers, C., Lalevic, D., Zafar, H. M., Cook, T.S. : Towardcompletestructured information extraction fromradiology reports using machine learning.Journal of digital imaging, 32,pp 554-564,2019.

[3] Krieger, F., Drews, P., & Funk, B.: Automatedinvoiceprocessing: Machine learning-based information extraction for long tailsuppliers. Intelligent Systemswith Applications, 20, 200285,2023.

[4] Fifita, F., Smith, J., Hanzsek-Brill, M. B., Li, X., & Zhou, M. ; Machine learning-based identifications of COVID-19 fake news usingbiomedical information extraction.Big Data and Cognitive Computing, 7(1), 46,2023.

[5] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ...& Jain, A. : Structured information extraction fromscientifictextwith large languagemodels.Nature Communications, 15(1), 141,2024.

[6] Luo, S., Yu, J. : ESGNet: A multimodal network model incorporating entity semantic graphs for information extraction from Chinese resumes. Information Processing& Management, 61(1), 103524, 2024.

[7] Hadji, A., Kholladi, M. K.  Automatic Opinion Extraction from Football-Related Social Media: A Gazetteer and Rule-Based Approach. NCAIA’2023, 61., 2023.

[8] Hadji, A., Kholladi, M. K., & Borisova, N. : Enhancing Spatial Information Extraction from Arabic Text: A Hybrid Approach with Ontology and Rule-Based. Ingénierie des Systèmes d'Information, 29(4),1261-1273 ,2024.

[9] Hadji, A., & Kholladi, M.K. : Advanced NPL Methods for Disaster Information Extraction: Analyzing JAPE Rules, Ontologies, and Machines Learning Approaches. Proceedings of the 3rd International Conference on computer Science’s Complex System and their Application CCSA’2024, Computer Science book series , Springer Nature. 2024. (In press)

[10] Kurani, A., Doshi, P., Vakharia, A., & Shah, M . A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 10(1), 183-208, 2023.

[11] Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R.  Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5), 1774-1785,2017.

[12] Lee, C. S., Cheang, P. Y. S., & Moslehpour, M.  Predictive analytics in business  analytics: decision tree. Advances in Decision Sciences, 26(1), 1-29,2022.

[13] Gusnina, M., , Salamah, U Student Performance Prediction in Sebelas Maret University Based on the Random Forest Algorithm. Ingenierie des Systemes d'Information, 27(3), 495.2022.

[14] Gutierrez, F., Dou, D., Fickas, S., Lanka, S., Zong, H.: Hybrid Ontology-Based Information Extraction System. J. Inf. Sci. 42(6), 798–820 , 2016.

[15] Maynard, D., Peters, W., Li, Y.: Metrics for Evaluation of Ontology-Based Information Extraction. In: Proceedings of the 15th International Conference on World Wide Web, Workshop on Evaluation of Ontologies for the Web 2006.

# A system for detecting traffic objects and estimating their distance

Bouguerne Imen[1,*,†], Douaoui Rayane [2,†]

[1,2] *University of Computer Sciences, Chadli Bendjedid El-Tarf*

**Abstract**

In recent years, advancements in computer vision have driven the automotive industry to develop driver assistance systems capable of significantly reducing road accidents. This thesis introduces a model designed for detecting road traffic objects and accurately estimating their distances.

The proposed model identifies and classifies objects such as cars, trucks, and pedestrians located in front of the vehicle. Deep learning algorithms are employed to precisely determine the nature of each object. Subsequently, the model estimates both the distance between the camera and detected objects, as well as between pairs of objects. Integrating these insights enables drivers to maintain better focus on the road and gain a clearer perception of their surroundings.

The system utilizes a Convolutional Neural Network (CNN) based on the ResNet-50 architecture, augmented with additional convolutional layers to enhance detection performance. Training was conducted on the Pascal VOC 2012 dataset, achieving an accuracy of 95.21% and minimal classification loss using the Nadam optimizer. We compared our model with the Faster R-CNN model, and ours achieved better performance.

**Keywords**

Computer vision, Deep Learning, Distance estimation, Object detection, ResNet-50.

## 1. Introduction

Technological developments in the last two decades have significantly facilitated access to digital systems in our daily lives. Among the key elements of digital systems, great attention is being paid to images. Currently, the representation and processing of digital images is the subject of very active research. The processing of images is a very large area that has evolved considerably over the past few decades.

One of the most powerful and convincing kinds of artificial intelligence is computer vision. Computer vision is the field in computer science that focuses on replicating some parts of the complexity of the human vision system and allowing computers to identify and process objects in images in the same way that person do. Until recently, computer vision was functioning on a limited basis. Thanks to advances in artificial intelligence and innovations in deep learning, significant progress has been made in recent years, surpassing human abilities in some detection and identification tasks.

Moreover, one of the most important applications of computer vision is detecting objects on the road, which contributes to advanced driving by tailoring, automating, and enhancing cars to increase safety and improve the driving experience. This technology is designed to alert drivers to potential risks and assist them in maintaining control of their vehicles to prevent or minimize accidents, as most accidents are caused by human errors. For this reason, systems also calculate

[*] Corresponding author.

[†] These authors contributed equally.

bouguerne-imen@univ-eltarf.dz (I. Bouguerne);

distances between the car and detected objects on the road, and between pairs of detected objects, aiming to provide users with a smart, safe, and comfortable driving experience.

The objectives of this thesis are as follows:
1. Identify and categorize road traffic objects and calculate the distances between detected objects and the camera, as well as between pairs of detected objects.
3. Develop an efficient Convolutional Neural Network (CNN) for these tasks.
4. Propose an approximation algorithm for distance computation.
5. Evaluate the performance of the proposed model in terms of classification accuracy, classification loss, model accuracy, Mean Squared Error (MSE), and Intersection over Union.

## 2. Related work

The "traditional object detection period (before 2014)" and the "deep learning-based detection period (after 2014)" are the two historical periods in which object detection has mostly developed. As noteworthy methods have developed after the introduction of deep learning to the field, we will examine contemporary approaches, which may be broadly divided into two categories: one-stage detection algorithms and two-stage detection algorithms.

**Table 1**
Road traffic objects detection methods.

| Ref. No | Year | Methods | | |
|---|---|---|---|---|
| | | Method used | | Results, Avantage, Limits, Robustness, …etc. |
| Algorithms based on ***C***onvolutional ***N***eural ***N***etworks | | | | |
| [3] | 2014 | ***Two-stage detection algorithm*** | R-CNN | ▪ mAP: 31.4% on ILSVRC 2013 and 62.9% on PASCAL VC 2010.<br>▪ Notable advancement over conventional techniques.<br>▪ Effective in employing selective search to produce abo 2000 ROIs.<br>▪ Long computation times because of several distin phases.<br>▪ Necessitates producing a lot of ROIs.<br>▪ Repeated computations result from overlapping regi suggestions.<br>▪ Less effective but more accurate than sliding windo techniques. |
| [3]<br>[4] | 2015 | | FAST R-CNN | ▪ It has a better detection quality than R-CNN and SPPne<br>▪ It uses a multi-task loss during the single stage of trainin<br>▪ The network's layers can all be updated through trainin<br>▪ Selective search is sluggish, resulting in a high processi time; it does not require disk storage. |
| [3] | 2017 | | FPN | ▪ It functions as a feature extractor and can alter th extractors of other detectors.<br>▪ FPN presents a pyramid approach to improve the quali of features for object detection.<br>▪ The architecture used by FPN is layered, with increasi semantic values at each step. |

| | | | | |
|---|---|---|---|---|
| | | | | ▪ The efficacy of FPN is mostly reliant on the semant layers; high-resolution layers are generated by leveragi layers with rich semantics.<br>▪ Effective connection management is one of FPN challenges.<br>▪ At every level, rich semantic information is available.<br>▪ Cutting off top-down links results in a reduction precision. |
| [5] | 2018 | | CORNERNET | ▪ The innovative method achieves competitive results l using key point detection (upper left and lower rig corners) for object detection.<br>▪ Absence of anchors lowers computational overhead ar simplifies detection.<br>▪ Flexibility: Enables customized feature extractic networks since there are no pre-trained models ar training begins from scratch.<br>▪ Corner Pooling: Presents a novel pooling technique th raises the precision of corner key point identification.<br>▪ Training Complexity: Since training is done from scratc it is more time- and computationally-intensive.<br>▪ Edge Cases: Possible issues with partially concealed poorly defined corners of objects.<br>▪ Generalization: Good generalization across datase nonetheless, the complexity of the scene and the objec affects performance.<br>▪ Scalability: For consistent performance, more testing c bigger, more varied datasets is necessary.<br>▪ Future study will be influenced by the switch fro anchor-based to key point detection in key poi detection.<br>▪ Corner Pooling: Increases precision |
| [6] | / | | YOLOP | ▪ YOLOP is an extremely effective multi-task netwo created to handle the three crucial tasks of autonomo driving: lane identification, segmentation of drivab zones, and object detection. It is noteworthy because it the first system to accomplish real-time performance c embedded devices.<br>▪ When YOLOv5s, MultiNet, DLT-Net, and Faster R-CN findings are compared, YOLOP produces superi outcomes. |

## 3. Proposed approach

Our system detects traffic objects (car, truck, pedestrian, bicycle, traffic light, motorcycle, bus, stop sign, etc.) and measure the distance between camera and object and between pair of detected objects in the images. Our system begging with upload the dataset then it prepare the data to work with. The dataset is splited to train set and test set. After that we trained the model with those data, the model will be predict this part of detection is ready.

The second part is estimate distance; we will take the results to make them as an input to our distance estimation algorithm. With the interface we will see carefully the results of the system.

We represent below the general architecture of our system.

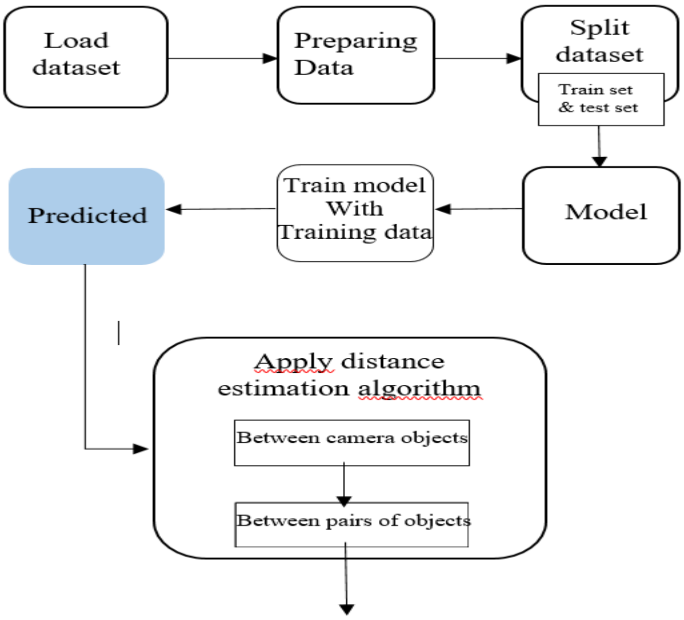


**Figure 1:** Our system architecture

We utilize ResNet50 for several purposes, particularly in the context of object identification and classification tasks:

**Rich architecture**: ResNet50, a 50-layer deep convolutional neural network, is very helpful for difficult tasks like object detection and classification since it can extract complicated features from the data.[7]

**Residual Learning**: The introduction of residual connections, also known as skip connections, is one of the main advances of ResNet50. By reducing the effects of the vanishing gradient issue, these connections enable the network to train more efficiently even at deeper depths.[7]

**Performance**: ResNet50 is a dependable option for many computer vision jobs because of its outstanding performance on multiple benchmarks. Its well-balanced architecture offers a decent compromise between computing efficiency and precision.

**Versatility**: ResNet50 is adaptable and may be used for a wide range of tasks, including segmentation and object detection, in addition to image classification. By building bespoke layers on top of the underlying model, ResNet50's feature extraction capabilities can be used for a variety of downstream applications.[7]

In our model, ResNet50 is used as the basic model for feature extraction:

High-level features are extracted from the input photos using the pre-trained ResNet50 model. This facilitates the use of the potent representations that ResNet50 has acquired.

In order to work with ResNet50 we need to modify in the architecture we have two parts the first one about the main ResNet50 and the second one about adding convolutional layers, in this part will explain the first one, the next title for the second one.

After importing ResNet50, we excluded the fully connected layers at the top, which are typically used for classification, which allow us to add our own layers for custom tasks (object detection), from the convolutional layers we need the output feature maps, which represents in 7*7*2048 dimensional array.

These feature maps 7*7*2048 serve as the foundation for predicting bounding box coordinates and object classes. The modifications shown below:

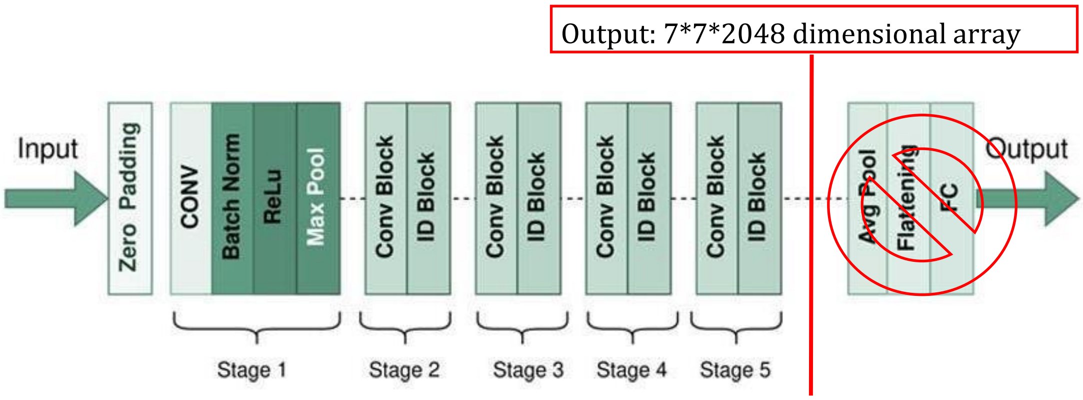


**Figure 2 :** Our system architecture

After that we Freeze Layers of the pre-trained ResNet50 model. This prevents their weights from being updated during training, the raisons why to freeze those layers are:

Avoid Overfitting: Assists in preserving stable representations of previously trained layers, thereby lowering the possibility of overfitting in the case of insufficient training data.

Computational efficiency: Lowers the amount of memory needed and training times by reducing the number of parameters that need to be updated.

Early Stages of Transfer Learning: When fine-tuning or transfer learning is just starting, layers are usually frozen. As a result, you may use the pre-trained layers for feature extraction during the early training epochs and concentrate on changing only the weights of the newly added layers.

### 3.1. Description of the architecture

We have developed a model architecture based on the ResNet50 framework, specifically intended for intricate image processing applications like object detection. Images with a resolution of 224x224 and three color channels (RGB) are initially accepted by the input layer. The crucial layer, conv5, is obtained after four convolutional layers. It employs 512 and 2048 filters to reduce spatial dimensions to 7x7 and to increase depth to 2048. The 7x7x2048 dimensional array that represents the final feature map from the ResNet50 backbone provides the input for later layers that are designed for certain tasks. The architecture has two Conv2D layers after ResNet50 . The depth is first lowered from 2048 to 512 and then again to 256 in the second. Two independent output heads intended for bounding box prediction and classification tasks subsequently process these improved feature maps. In order to represent bounding box coordinates, the output is reshaped to (49, 4) by the bounding box output head using a Conv2D layer with four filters. Class probabilities are represented by reshaping the output to (49, 10) using a Conv2D layer with 10 filters in the classification output head. This architecture is ideal for object detection since it makes use of the strong feature extraction capabilities of ResNet50 and the adaptability of extra Conv2D layers. With the effective reduction of spatial dimensions and enhancement of feature depth, it guarantees precise and comprehensive predictions for bounding boxes and class labels. Remaining blocks and task-specific layers together provide a strong design that can handle challenging image analysis tasks.

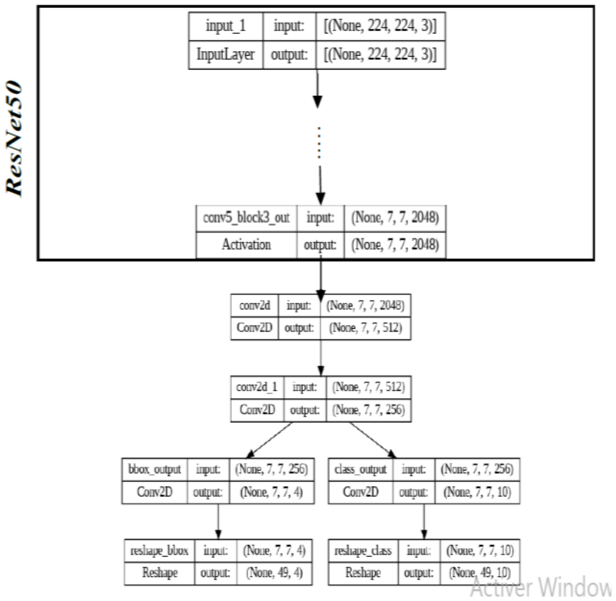

ResNet50
input_1 | input: | [(None, 224, 224, 3)]
InputLayer | output: | [(None, 224, 224, 3)]
conv5_block3_out | input: | (None, 7, 7, 2048)
Activation | output: | (None, 7, 7, 2048)
conv2d | input: | (None, 7, 7, 2048)
Conv2D | output: | (None, 7, 7, 512)
conv2d_1 | input: | (None, 7, 7, 512)
Conv2D | output: | (None, 7, 7, 256)
bbox_output | input: | (None, 7, 7, 256)
Conv2D | output: | (None, 7, 7, 4)
class_output | input: | (None, 7, 7, 256)
Conv2D | output: | (None, 7, 7, 10)
reshape_bbox | input: | (None, 7, 7, 4)
Reshape | output: | (None, 49, 4)
reshape_class | input: | (None, 7, 7, 10)
Reshape | output: | (None, 49, 10)

**Figure 3 :** Our system architecture

### 3.2. Dataset

The dataset consists of 20 classes. 11,530 photos with 27,450 annotations for regions of interest make up the dataset, which can be used for training and validation. The dataset has twenty classes, which are as follows:

Person: person.

Animal: Sheep, dogs, horses, cats, birds, and cows.

Vehicles: motorbike, train, bus, bicycle, boat, airplane, and car.

Indoors: TV/monitor, sofa, dining table, bottle, and potted plant.

Trainval (2913 images), train (1464 images), test (1456 images), and val (1449 images) are the three subsets that make up the dataset. For development tasks like feature selection and parameter adjustment, the trainval set is employed. Results on unseen data are reported using the test set. It is crucial to remember that algorithms must only be executed once on the test data in order to guard against misuse of the evaluation server.

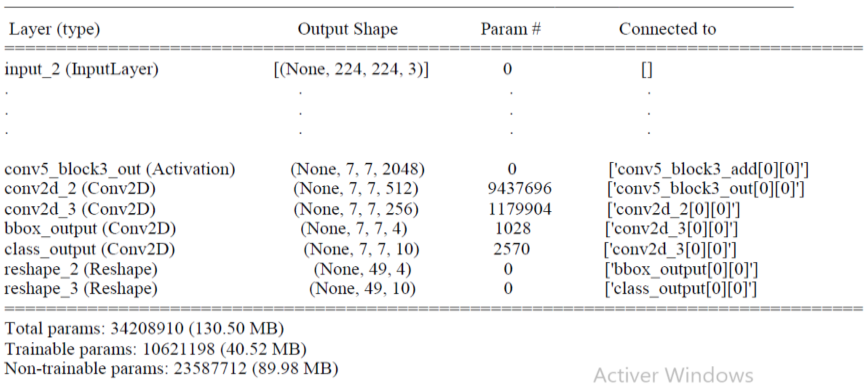
The most common problems are segmentation (which class does each pixel belong to), detection (where are the examples of a certain object class in the image? ), and classification (does the image contain any instances of a given object class?). Two more issues exist in addition to these: action classification (i.e., what action is the person indicated in this image performing?) and person arrangement (where are the people's hands, feet, and heads in this picture?).

Because of its size and excellent annotations, the PASCAL VOC 2012 dataset has been extensively utilized as a benchmark for computer vision applications. It's also important to note that this dataset contains the neutral class. This unique class permits less accurate manual tagging of object boundaries during neural network training. However, with the advent of interactive systems that offer high-quality pixel-level segmentation and have substantially sped up human labeling, this strategy is becoming less common.[8]

### 3.3. Model configuration and number of parameters

Our model is intended for item recognition and categorization in image processes input photos with three color channels (RGB) and a size of 224 x 224. It seems to be intended for object detection jobs. The input layer of the model is the first layer that feeds into the convolutional layers of the ResNet50 backbone network, which has already been trained. High-level characteristics are extracted using this backbone, which shrinks the spatial dimensions to 7x7 with 2048 channels. Further convolutional layers analyze these features after the backbone, lowering the channel dimensions successively to 512 and subsequently to 256. Next, two concurrent convolutional layers are applied to the generated feature maps: one predicts bounding box coordinates using four channels, and the other predicts class probabilities using ten channels, which correspond to ten distinct object classes. In order to prepare these outputs for additional processing or loss calculation, they are reshaped to 49x4 and 49x10, respectively. 34,208,910 parameters make up the model; 10,621,198 of those are trainable, while 23,587,712 are not. This architecture is appropriate for object detection applications because it makes use of the feature extraction capabilities of a pre-trained backbone while tailoring the output layers for the particular tasks of localization and classification.

Model:"model_1"



```
_________________________________________________________________________________
 Layer (type)                  Output Shape          Param #     Connected to
=================================================================================
input_2 (InputLayer)          [(None, 224, 224, 3)]     0              []
.                                  .                    .              .
.                                  .                    .              .
.                                  .                    .              .

conv5_block3_out (Activation)   (None, 7, 7, 2048)      0         ['conv5_block3_add[0][0]']
conv2d_2 (Conv2D)               (None, 7, 7, 512)     9437696     ['conv5_block3_out[0][0]']
conv2d_3 (Conv2D)               (None, 7, 7, 256)     1179904     ['conv2d_2[0][0]']
bbox_output (Conv2D)            (None, 7, 7, 4)        1028       ['conv2d_3[0][0]']
class_output (Conv2D)           (None, 7, 7, 10)       2570       ['conv2d_3[0][0]']
reshape_2 (Reshape)              (None, 49, 4)          0         ['bbox_output[0][0]']
reshape_3 (Reshape)              (None, 49, 10)         0         ['class_output[0][0]']
=================================================================================
Total params: 34208910 (130.50 MB)
Trainable params: 10621198 (40.52 MB)
Non-trainable params: 23587712 (89.98 MB)
```

**Figure 4 :** Our system architectur

## 4. Experiments, discussion the obtained results

In this section, we will experiment with several optimization techniques and to further enhance the performance of our model.

Great to hear that our model achieved a high accuracy of 0.9518 in Epoch 36 with a low loss of 0.2170 in the training set. Furthermore, our validation accuracy is also high at 0.9499 with a low validation loss of 0.2415, indicating that our model is performing well on both the training and validation data. This is a good sign that our model is generalizing well to unseen data and can be used for predicting new data. The classification model accuracy and validation accuracy are similarly high at 0.9518 and 0.9499, respectively. The Mean Squared Error (MSE) and validation MSE are low at 0.0351 and 0.0381, respectively, while the Intersection over Union (IoU) and validation IoU are also high at 0.8611 and 0.8590, respectively, further confirming the model's robustness.

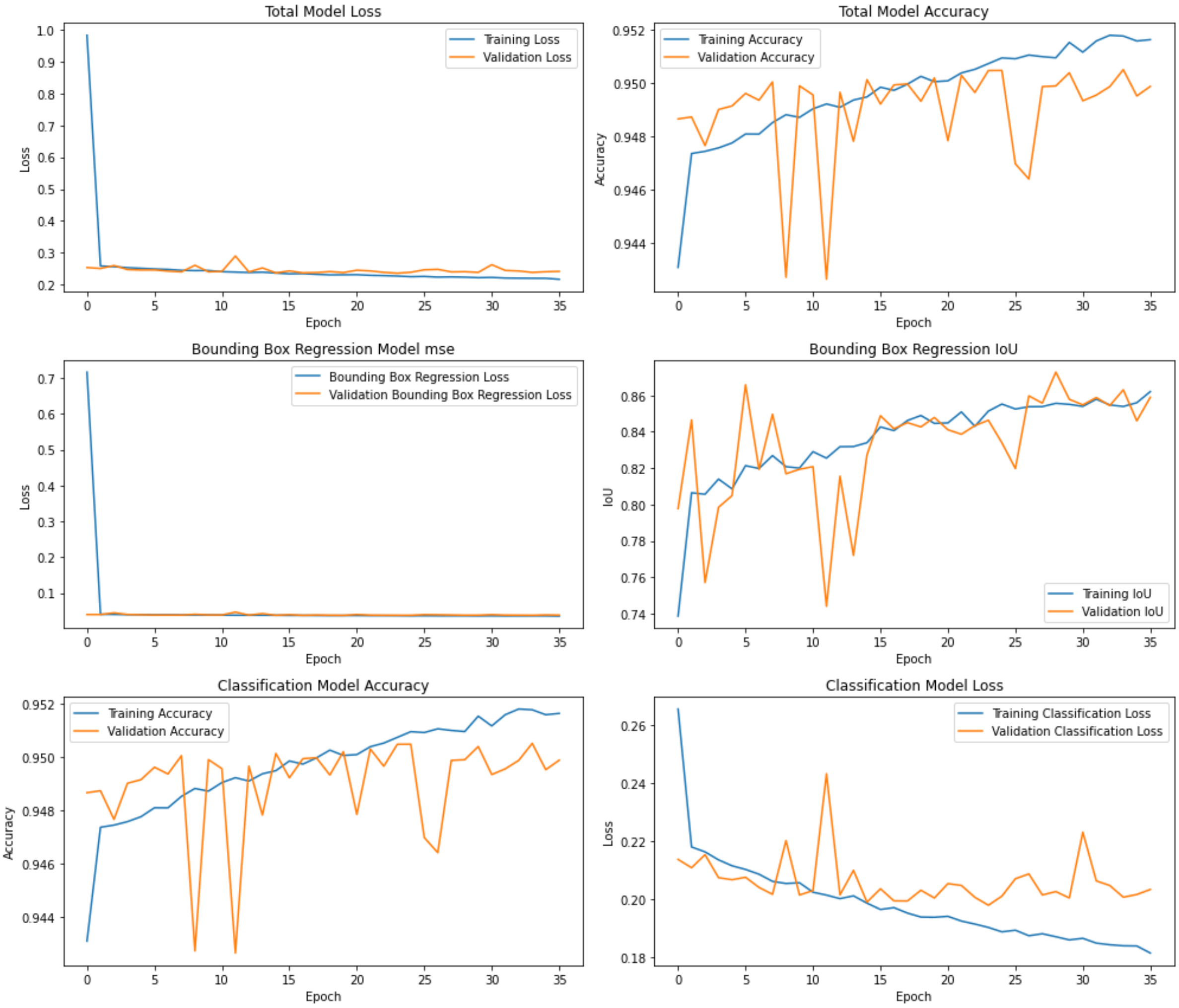
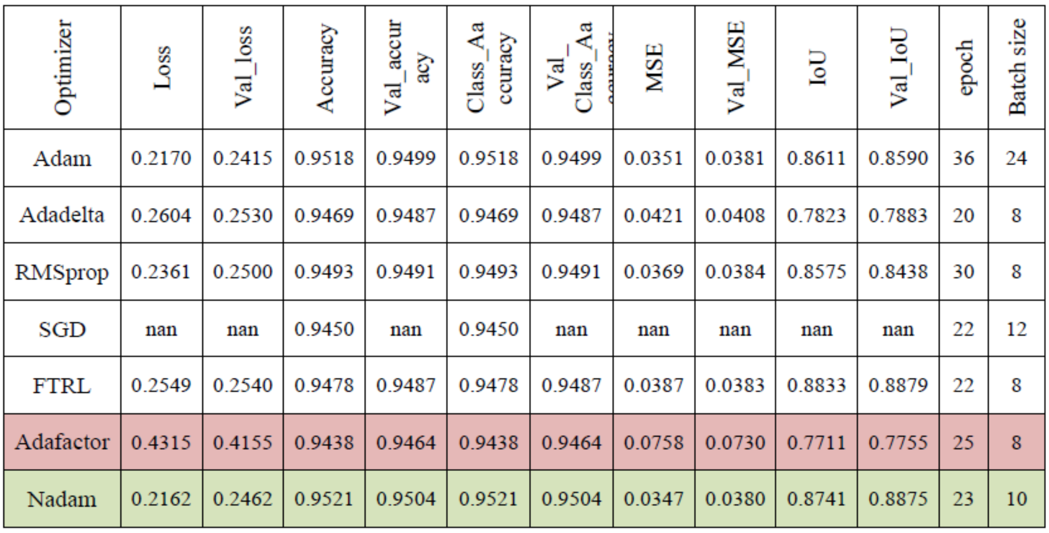


**Figure 5 :** Charts of adem's results.

**Table 2**

Comparison of different optimizers.



| Optimizer | Loss | Val_loss | Accuracy | Val_accuracy | Class_Aaccuracy | Val_ Class_Aaccuracy | MSE | Val_MSE | IoU | Val_IoU | epoch | Batch size |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Adam | 0.2170 | 0.2415 | 0.9518 | 0.9499 | 0.9518 | 0.9499 | 0.0351 | 0.0381 | 0.8611 | 0.8590 | 36 | 24 |
| Adadelta | 0.2604 | 0.2530 | 0.9469 | 0.9487 | 0.9469 | 0.9487 | 0.0421 | 0.0408 | 0.7823 | 0.7883 | 20 | 8 |
| RMSprop | 0.2361 | 0.2500 | 0.9493 | 0.9491 | 0.9493 | 0.9491 | 0.0369 | 0.0384 | 0.8575 | 0.8438 | 30 | 8 |
| SGD | nan | nan | 0.9450 | nan | 0.9450 | nan | nan | nan | nan | nan | 22 | 12 |
| FTRL | 0.2549 | 0.2540 | 0.9478 | 0.9487 | 0.9478 | 0.9487 | 0.0387 | 0.0383 | 0.8833 | 0.8879 | 22 | 8 |
| Adafactor | 0.4315 | 0.4155 | 0.9438 | 0.9464 | 0.9438 | 0.9464 | 0.0758 | 0.0730 | 0.7711 | 0.7755 | 25 | 8 |
| Nadam | 0.2162 | 0.2462 | 0.9521 | 0.9504 | 0.9521 | 0.9504 | 0.0347 | 0.0380 | 0.8741 | 0.8875 | 23 | 10 |

This table shows the performance metrics (accuracy and loss) for different optimization algorithms along with the corresponding validation metrics and the number of epochs. The best-performing algorithm based on validation accuracy and loss is highlighted in green, while the worst-performing algorithm is highlighted in red. Based on the provided results, it seems like the best optimizer (model) used is Nadam. It has a high accuracy of 0.9521 and a low loss of 0.2162 on the training data, as well as a high validation accuracy of 0.9504 and a low validation loss of 0.2462. Additionally, it achieved these results in only 23 epochs.

Based on these results, the **Nadam optimizer** stands out as the most effective, achieving the highest accuracy and the lowest loss in both training and validation datasets, demonstrating excellent generalization and performance. Conversely, the **Adafactor optimizer** shows the poorest performance, with the highest loss and the lowest accuracy, indicating suboptimal model training and validation results.

The system's classification accuracy, when using the Nadam optimizer, was determined to be 95.21%. This high accuracy indicates the model's effectiveness in correctly identifying and categorizing various traffic objects such as cars, pedestrians, and bicycles.

In addition to classification accuracy, the system's low classification loss and MSE, coupled with a high IoU, suggest that the model is not only accurate in detecting objects but also reliable in localizing objects. The Nadam optimizer's performance highlights its suitability for this specific application, making it the preferred choice for optimizing the traffic object detection.

## 5. Conclusion

In order to improve road traffic safety, this thesis investigated the use of computer vision and image processing algorithms for the identification and measurement of distances between different objects, such as vehicles, animals, and bicycles. This work's larger background is the growing dependence on artificial intelligence (AI) and digital technologies in daily life, especially when it comes to improving driving safety using cutting-edge detection systems. The project specifically sought to create an approximation method and a Convolutional Neural Network (CNN) based on the ResNet-50 architecture for precise distance computation.

Several challenges were encountered during this research. The first challenge was conceptualizing the best implementation idea, which took considerable time and effort. The second challenge involved finding a suitable dataset. Initially, the MSCOCO dataset, with over 80 classes, was considered ideal for object detection. However, its size of 26GB posed a barrier, making it impractical to train on a modest laptop. The third challenge was training the model: with 34 million parameters and a 4GB dataset, training for 20 epochs took over 6 hours, which was resource-intensive.

The solution involved successfully developing a robust CNN model based on ResNet-50 for object detection and a novel algorithm for distance estimation. The system was rigorously evaluated using metrics such as classification accuracy, classification loss, model accuracy, Mean Squared Error (MSE), and Intersection over Union (IoU). The results demonstrated that the system could accurately detect and classify road traffic objects and compute distances with high precision, outperforming some existing models like Faster R-CNN.

## References

[1] Anik Datta, Tamara Islam Meghla, Tania Khatun. Road Object Detection in Bangladesh using Faster R-CNN: A Deep Learning Approach. 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON- ECE). 12 April 12th, 2021 ieee.

[2] Yandan Kong, Kai Liu, Zhihong Liang, Tiancheng Liu. Research on small object detection methods based on deep learning. 2022 IEEE 4th International Conference on Power, Intelligent Computing and Systems (ICPICS). September 07th, 2022 ieee.

[3] C. Bhagya, A. Shyna . An Overview of Deep Learning Based Object Detection Techniques. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT) . June 21th, 2019 ieee.

[4] [11] Jun Deng,a, Xiaojing Xuan, Weifeng Wang, Zhao, Hanwen , Zhiqiang,Wang. A review of research on object detection based on deep learning.

[5] https://www.scirp.org/journal/paperinformation?paperid=115011 27/01/2024

[6] Dong Wu Manwen Liao Weitian Zhang Xinggang Wang Xiang Bai Wenqing Cheng Wenyu Liu. YOLOP: You Only Look Once for Panoptic Driving Perception. arXiv:2108.11250

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

[8] Mark Everingham · S. M. Ali Eslami · Luc Van Gool · Christopher K. I. Williams · John Winn · Andrew Zisserman. The PASCAL Visual Object Classes Challenge: A Retrospective. Springer Science +Business Media New York 2014. 25 June 2014

[9] Chollet, F. and Team, K. (2021). Keras documentation: Optimizers. https://keras.io/api/ optimizers. Accessed on May 7th, 2023

[10] Yanying Zhang, Xiaolin Zheng. Development of Image Processing Based on Deep, Learning Algorithm. 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). ieee may 23th 2022.

[11] M Sandeli, « traitement d'images par des approches bio-inspirées application à la segmentation d’images », Constantine 2 university. 2014.

# Advancements in Suicide Ideation Detection: A Comprehensive Literature Review

Ferdaous Benrouba[1,*,†], Rachid Boudour[2,**,‡]

[1] *Chadli-Bendjedid University of Eltaref*
[2] *Badji Mokhtar University of Annaba*

**Abstract**

This comprehensive literature review explores the significant advances made in the field of suicide identification using various advanced algorithms and techniques. Natural language processing (NLP), machine learning (ML), and deep learning (DL) methods have proved invaluable for identifying and understanding suicidal ideas in various data sources, including social media content and clinical notes. Several studies have highlighted the potential of interdisciplinary approaches that combine NLP's language analysis capabilities, ML's predictive strength, and DL's subtle understanding of complex text data. The results of these studies demonstrate the high precision achieved by different models and show their effectiveness in identifying early signs of suicide thoughts. These progresses offer significant promises for enhancing suicide prevention efforts and early intervention strategies, ultimately contributing to improving the results of mental health for people at risk. However, ongoing research, larger datasets, and algorithm improvements are essential to ensure their ongoing effectiveness and practical application in suicide prevention and mental health support.

**Keywords**

Machine learning, Natural Language Processiong, Deep Learning, Artificial I ntelligence

## 1. Introduction

This in-depth study delves into the forefront of research and methodologies employed in the critical domain of suicide ideation detection. Against the backdrop of alarming statistics from the World Health Organization (WHO)[1], which reveal that every year 703,000 people take their own lives, with countless others attempting suicide, this study takes a closer look at how cutting-edge techniques in Natural Language Processing, Machine Learning, and Deep Learning are harnessed to address this pressing issue. The WHO statistics highlight the profound impact of suicide, extending far beyond the individual act itself. Each suicide is a tragedy that reverberates through families, communities, and entire nations, leaving a lasting scar on those left behind. Moreover, suicide does not discriminate based on age, occurring throughout the lifespan and tragically becoming the fourth leading cause of death among 15–29-year-olds globally in 2019. In the face of these sobering facts, this literature review embarks on an exploration of a multitude of studies that endeavor to identify suicidal ideation across various data sources, including social media content and clinical notes. By providing a detailed synthesis of these studies, we aim to shed light on the advancements and challenges within this field. Fellow, as we delve into the introduction section to present a holistic view of how Artificial Intelligence and its subfields play a pivotal role in detecting suicidal ideation, we will commence by providing concise definitions for three fundamental components: Machine Learning, Deep Learning, and Natural Language Processing. To enhance clarity, each definition will be accompanied by an illustrative schema depicting the core principles of these domains.

*icasa24 : 2nd International Conference on Autonomous Systems and their Applications.*

[*] Corresponding author.

[†] These authors contributed equally.

f.benrouba@univ-eltarf.dz (F. Benrouba); racboudour@yahoo.fr (R. Boudour)
https://github.com/benrouba (F. Benrouba)
0000-0002-3853-8426 (F. Benrouba); 0000-0002-5060-1694 (R. Boudour)

### 1.1. Machine Learning

This paper will use the definition proposed by Yu et He, which denotes the procedure by which a computer system acquires the ability to learn from data and enhance its performance on a specific task without direct programming for that task [2]. This definition is consistent with the widely acknowledged concept that machine learning involves the utilization of algorithms and statistical models, enabling computer systems to iteratively enhance their performance on a designated task through learning from data, without explicit programming. Machine learning encompasses diverse techniques, including classification, combination methods, and learning to rank, all integral components in facilitating machines to learn from data and make informed decisions based on that acquired knowledge. Furthermore, the integration of machine learning principles into various disciplines, such as medical education and economics, underscores the extensive applicability and significance of machine learning across diverse domains [3]. In the machine learning process, the first step involves defining the objective, followed by the collection of relevant data. After data preprocessing, including cleaning and feature engineering, a suitable machine learning algorithm is selected based on the problem's nature. The model is then trained using a portion of the data, evaluated for performance, and fine-tuned through hyperparameter adjustments. Once optimized, the model is deployed for predictions on new data, with ongoing monitoring and maintenance to ensure continued effectiveness. The process is illustrated in the Figure 1, depicting key stages from defining the problem to deploying and maintaining the trained model.

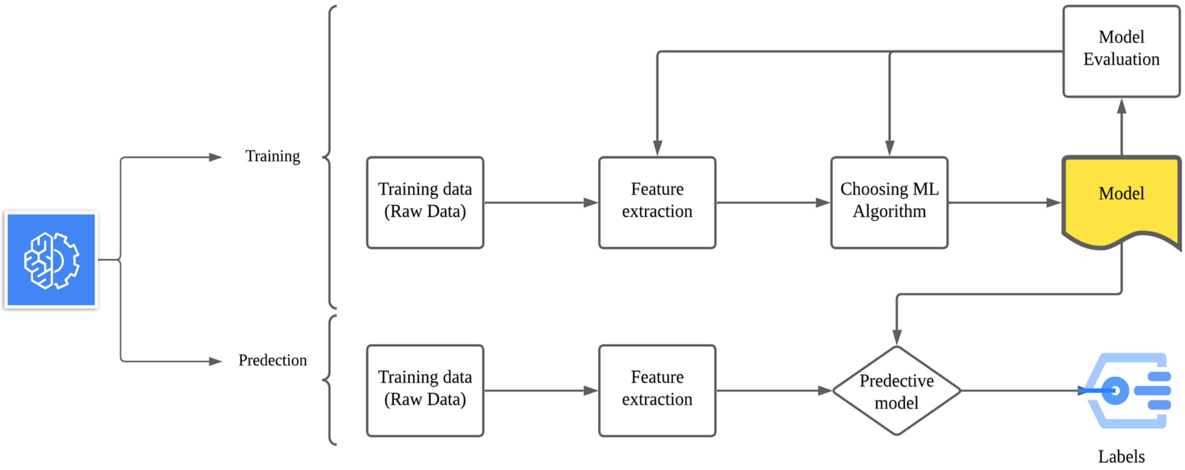


**Figure 1:** Machine learning workflow

### 1.2. Deep Learning

Deep learning, a potent technique within the realm of machine learning, has garnered considerable attention across diverse domains due to its proficiency in handling extensive and intricate datasets, enabling the discernment of novel features and patterns. It employs neural networks with multiple layers for hierarchical feature learning and pattern classification [4].(refer to Figure 2 for visual representation). The versatility of deep learning is evident in its successful applications spanning computer vision, genomics, proteomics, and medical image analysis [5]–[8]. In brain MRI segmentation, deep learning surpasses the limitations of classical machine learning algorithms, proving instrumental in identifying new imaging features for quantitative analysis. Its impact extends to clinical genomics, where deep learning algorithms process vast and complex genomic datasets, underscoring its significance in the medical arena.

The influence of deep learning reaches beyond conventional scientific fields, revolutionizing domains like bioinformatics, ecology, and education. In education, deep learning entails critical thinking, the

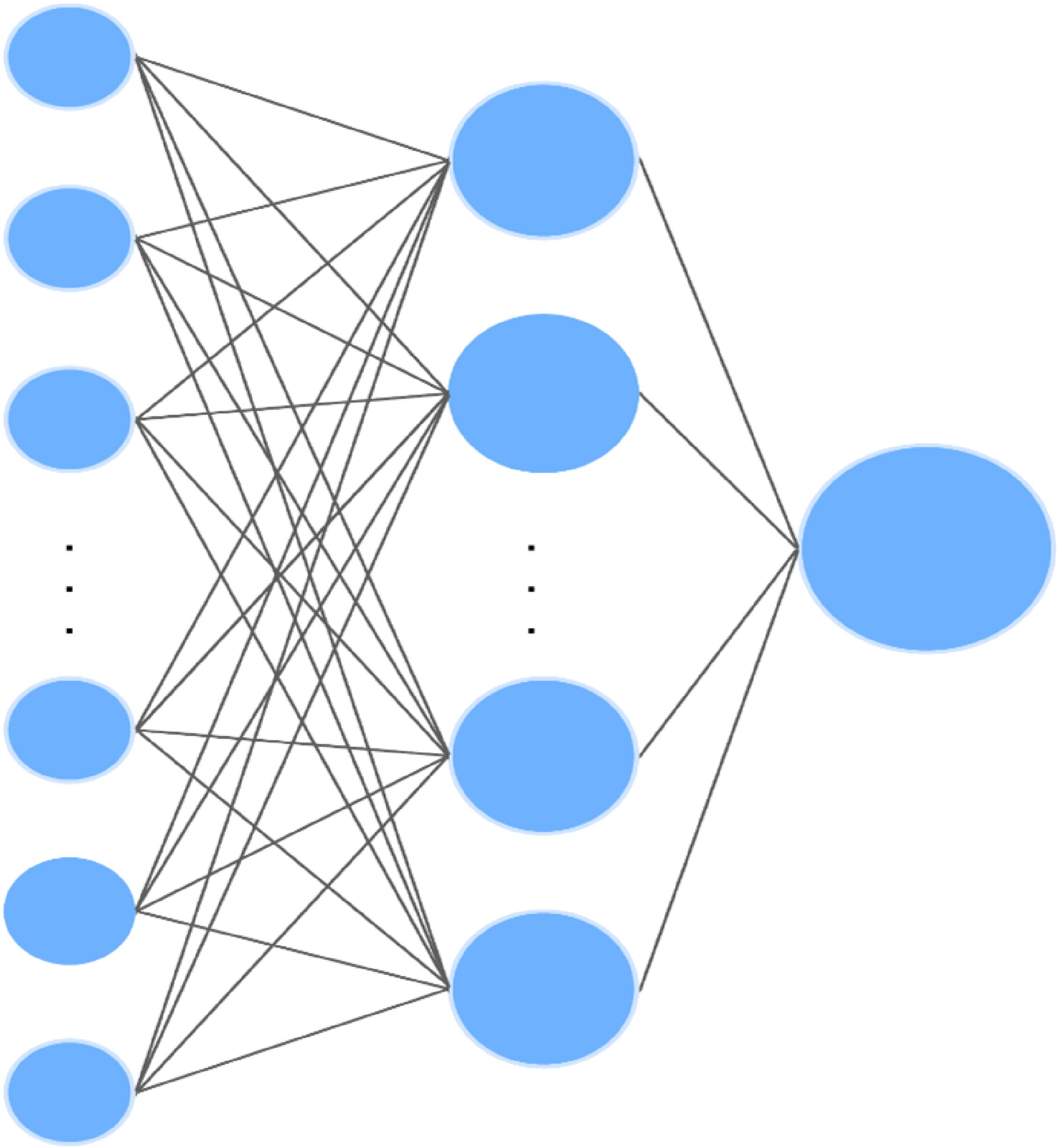


**Figure 2:** Illustration of the neural networks explained in the context of deep learning models.

integration of new knowledge with existing understanding, and the formation of novel connections and concepts. Moreover, its application in big data analysis and biometrics emphasizes its suitability for managing substantial data volumes from diverse sources.

### 1.3. Natural Language Processing

Natural Language Processing is a discipline employing computational methods for the examination and generation of human language content [9]. NLP systems are designed to comprehend how textual and spoken materials are processed by computerized systems and operated on computing devices [10]. Recent research emphasizes the perpetuation of biases within NLP systems, particularly evident in machine translation and dialogue systems [11]. Ethical considerations surrounding NLP have garnered increased attention, with investigations delving into concerns such as predictability, privacy, decision-making, responsibility, and bias [12]. Furthermore, the influential capabilities of NLP techniques have impacted various facets of daily life, presenting ethical and social challenges, including potential

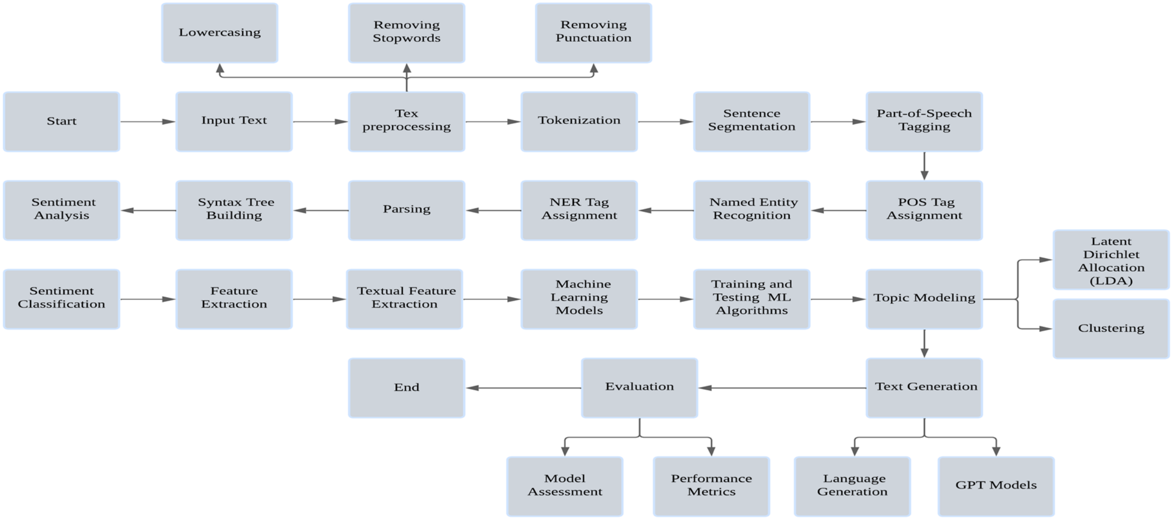


**Figure 3:** Comprehensive Natural Language Processing (NLP) Workflow: From Text Input to Model Evaluation.

harm from language models [13]. The exploration of the interplay between native human languages and computer systems is encapsulated by the field of NLP [14]. From the utilization of deep neural networks to the integration of multi-modal features and NLP-driven linguistic analysis, this study offers a comprehensive overview of the diverse approaches used to tackle this sensitive issue. It is our hope that through the critical evaluation of results and discussions, we can underscore the potential of these methods in early intervention and suicide prevention. Ultimately, this research seeks to make a meaningful contribution to the ongoing efforts to address mental health concerns in the digital age. This paper is organized as follows: in Section 2, we introduce the different methods, techniques, and algorithms used to detect suicidal ideation; Section 3 presents the results and discussion; and finally, in Section 4, we provide a conclusion to this work.

## 2. Methodologies

### 2.1. Machine Learning Techniques for Suicidal Ideation Detection

In recent years, there has been a burgeoning interest in using machine learning techniques to detect signs of thoughts at an early stage. Researchers have come up with methods that combine natural language processing and various machine learning algorithms to analyze communication and identify potentially suicidal content. In this section we will explore studies that utilize datasets and machine learning algorithms to predict and recognize suicidal thoughts. These studies provide insights, into how these techniques can contribute to suicide prevention efforts. For quick reference, Table 1 conveniently summarizes the findings from these studies for reference while Figure 5 illustrates the predominant architecture utilized across the majority of the researchs. In a research conducted by Lee et al [15], the authors employed various machine learning algorithms, including Logistic Regression, SVM, Random Forest (RF), and Extreme Gradient Boosting (XGBoost), to predict suicidal ideation and suicide planning or attempt in a Korean adult population. They utilized a dataset obtained from population survey data, incorporating a range of psychosocial features such as depressive symptoms, self-esteem, economic condition, and others as predictors. The algorithms achieved classification accuracies ranging from 80% to 90%, with XGBoost demonstrating the highest performance. This study demonstrates the potential of machine learning in identifying individuals at risk of suicidal ideation and planning or attempt while highlighting the relevance of specific psychosocial factors in suicide risk assessment. In their study titled "Assessment of Supervised Classifiers for Detecting Messages with Suicidal Ideation," [16] the authors employed 28 supervised classification algorithms, including RandomCommittee, RandomTree, and

Kstar, to detect suicidal ideation in text messages. They extracted linguistic and text-based features such as Part-of-Speech tags, word stems, synsets, lemma, and word frequency from the dataset, primarily the Life Corpus. This corpus contains text messages, including those with suicidal ideation, sourced from platforms like Twitter, and is openly available under a Creative Commons license. Performance was evaluated using traditional metrics like F-measure and ROC Area, with KStar using POS-SYNSET-NUM and POS-NUM features showing the best performance. Despite the small dataset size, the study achieved statistically significant results, emphasizing the potential of machine learning in suicide prevention and the need for ethical data collection when expanding datasets with social network data. The paper by Birjali et al [17] introduces a method for preventing suicide by analyzing sentiment in Twitter data through machine learning and semantic analysis. The study utilizes the Weka tool and a dataset comprising 892 tweets, dividing them into training and test sets. Various machine learning algorithms, such as IB1, J48, CART, SMO, and Naive Bayes, are applied to classify tweets into two categories: those with a risk of suicide and those without. The precision of these classifiers ranges from 61% to 89.5%, with SMO achieving the highest precision. The paper also presents a semantic similarity measure based on WordNet for analyzing the tweets. The method shows promise in identifying potentially suicidal content on Twitter, with implications for mental health monitoring and intervention. Authors In [18] conduct a comprehensive investigation into the use of machine learning for detecting suicidal ideation in user-generated text. The study employs three distinct algorithms—Support Vector Classification (SVC), Extra Trees, and Random Forest—and rigorously assesses their performance using statistical metrics. The results reveal varying degrees of precision, recall, and F-score for each algorithm, with SVC achieving an F-score of 0.90 (weighted), Extra Trees 0.90 (weighted), and Random Forest 0.90 (weighted). This analysis demonstrates the effectiveness of these models in accurately classifying text into positive and negative classes for suicidal ideation, highlighting their potential in mental health applications. These statistical findings contribute to the development of transparent and reliable AI tools for suicide risk assessment, paving the way for improved mental health support systems. In the upcoming subsection, we will explore studies centered around deep learning algorithms, a specialized sub-domain within the broader field of machine learning. This separation is intended to streamline our discussion and provide a more organized overview of the innovative approaches that leverage deep learning techniques for detecting suicidal ideation, you can refer to Figure 4 for a visual representation of the categorization of suicide ideation detection algorithms and datasets.

### 2.1.1. Deep Learning Techniques for Suicidal Ideation Detection

The studies presented in this subsection, which utilize deep learning algorithms and techniques, are summarized in Table 2. In their study on detecting suicidal ideation in social media posts, the authors in [19] utilized deep learning techniques, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and attention mechanisms, to analyze text data collected from Reddit, with a particular focus on the SuicideWatch forum. They preprocessed the textual content, applied word embeddings, and identified linguistic indicators of suicidal ideation. The ensemble model, combining CNNs, LSTMs, and attention mechanisms, achieved a high accuracy of 90.3% and an F1-score of 92.6% in classifying posts into categories like "No Risk," "Low Risk," "Moderate Risk," and "Severe Risk." The dataset used for training and evaluation included posts from Reddit users, with different labels reflecting the risk level of suicidal ideation in the content. Priyamvada et al. [20] conducted an extensive investigation titled “Stacked CNN - LSTM approach for prediction of suicidal ideation on social media” that delves into the realm of deep learning techniques for identifying suicidal ideation in social media content. Their method ingeniously combines Convolutional Neural Networks and Long Short-Term Memory networks, all within a stacked architecture enriched by Word2Vec word embeddings. The study meticulously compared various machine learning algorithms, with XGBoost demonstrating superiority in baseline assessments. Notably, their innovative Stacked CNN - 2 Layer LSTM model, fortified with Word2Vec embeddings, achieved remarkable accuracy, reaching an impressive 93.92%. This model surpassed the individual CNN and LSTM classifiers. The paper provides a robust foundation for leveraging deep learning to enhance the detection of suicidal ideation, particularly within Twitter

**Table 1**
Comparative Analysis of Suicidal Ideation Detection Studies Using Machine Learning

| Method | Study Focus | Dataset | ML Used | Key Findings |
|---|---|---|---|---|
| Population-Based Study (Lee et al [15]) | Predicting suicidal ideation and planning or attempt in a population | Korean adult population survey data | Logistic Regression, SVM, RF, XGBoost | Achieved classification accuracies ranging from 80% to 90%. Highlighted specific psychosocial factors in suicide risk assessment. |
| Text Message Analysis (Acuña et al[16]) | Detecting suicidal ideation in text messages | Life Corpus (text messages) | Random-Committee, RandomTree, Kstar | KStar using specific linguistic features showed the best performance. Emphasized ethical data collection for expanding datasets with social network data |
| Semantic Analysis (Birjali et al [17]) | Preventing suicide through sentiment analysis of Twitter data | Twitter data | IB1, J48, CART, SMO, Naive Bayes | Precision ranged from 61% to 89.5%. Applied semantic analysis based on WordNet. Showed potential for mental health monitoring and intervention. |
| User-Generated Text Analysis (De Oliveira et al[18]) | Detecting suicidal ideation in user-generated text | User-generated text | Support Vector Classification (SVC), | Achieved high F-scores: SVC - 0.90 (weighted), Extra Trees - 0.90 (weighted), Random Forest - 0.90 (weighted). Effective in classifying text for suicidal ideation. |

data. In their paper titled “Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation”[21], The authors presented a study that employs weak supervision and deep learning techniques to classify clinical notes for the presence of "current" suicidal ideation. The research utilizes a large dataset of clinical notes, including 13,426 notes from 456 patients. Various machine learning algorithms, including TF-IDF Logistic Regression, TF-IDF SVM, Random CNN, and Word2vec CNN, are evaluated for their performance in identifying suicidal ideation. The statistical results indicate that the deep learning methods achieved a high accuracy of 94% in classifying clinical text, identifying 42 additional encounters and 9 patients indicative of suicidal ideation, which were missed by conventional ICD-9/10 coding. The study highlights the potential of weakly supervised machine learning for improving suicide prediction models, despite the challenges posed by clinical interpretability and the need for external validation. In [22], Aldhyani et al presented a study on the early detection of suicidal ideation in social media content. The research employs a dataset from Reddit, consisting of posts with suicidal and non-suicidal content. The authors utilize machine learning and deep learning techniques, including a CNN–BiLSTM model, for text analysis. The study achieved a remarkable accuracy of 95% in identifying suicidal ideation, surpassing previous methods with a significant margin, as indicated in the comparative analysis with other studies. The paper also explores the use of LIWC features and word clouds for visualization and analysis of the textual corpus. This comprehensive approach highlights the potential for using advanced algorithms to detect early signs of suicidal thoughts in online social media, which could aid in suicide prevention efforts. The authors in [23] detailed a study on detecting suicidal ideation in social media forums using deep learning techniques. The method relies on convolutional neural networks and long short-term memory networks for text analysis. Several deep learning algorithms, including CNN-LSTM hybrids, are discussed for text classification. The study reports promising results in identifying suicidal ideation from social media text, with high precision, recall, and F1-score values. In particular, XGBoost outperforms traditional text classification methods when considering both combined and single features, except for Statistics. Notably, the LSTM deep learning classifier achieves the highest performance among baseline approaches, with an accuracy of 91.7% and an F1 score of 92.6%. Multiple datasets from diverse sources were used, highlighting the need for more annotated data and exploring correlations with external factors like weather and family environment to improve detection accuracy.

**Table 2**
Summary of Deep Learning-Based Approaches and Performance Metrics in Suicidal Ideation Prediction

| Study | Deep Learning Techniques Used | Dataset and Data Source | Statistical Results |
|---|---|---|---|
| Renjith et al [19] | CNNs, LSTMs, Attention Mechanisms | Text data from Reddit, with a focus on the SuicideWatch forum | Accuracy: 90.3%, F1-score: 92.6 |
| Priyamvada et al [20] | CNNs, LSTMs, Word2Vec Word Embeddings | Twitter data (specific details undisclosed) | Accuracy: 93.92% |
| Cusick et al [21] | TF-IDF Logistic Regression, TF-IDF SVM, Word2Vec CNN, etc. | Clinical notes dataset with 13,426 notes from 456 patients | Accuracy: 94% |
| Aldhyani et al [22] | CNN–BiLSTM Model | Reddit dataset with suicidal and non-suicidal content | Accuracy: 95% |
| Tadesse et al [23] | CNNs, LSTMs | Multiple datasets from diverse sources | LSTM Accuracy: 91.7%, LSTM F1-score: 92.6% |
| Gyllensten et al [24] | RNN, LSTM, C-LSTM | Twitter dataset with manually annotated tweets | C-LSTM Accuracy: 81.2% |
| Hu et al[25] | Deep Neural Network Models | Structural MRI data of 288 MDD patients categorized into SA, SI, NS | Accuracy over 70% in classification tasks |
| Shin et al [26] | Logistic Regression, Random Forest, CNN, etc. | Dataset by Korea Youth Policy Institute between 2017 and 2020 | CNN Accuracy: 88.3%, CNN Precision: 90.1% |

The study conducted in [24] explores the detection of suicidal ideation in social media, particularly on Twitter, using deep learning models. The authors first generated a lexicon of relevant terms, collected a dataset of tweets, and manually annotated them for suicidal intent. They employed three deep learning models, including RNN, LSTM, and C-LSTM, along with traditional baseline methods, to classify tweets. The results showed that the C-LSTM model outperformed the other methods, achieving an accuracy of 81.2% and demonstrating its effectiveness in identifying suicidal ideation in tweets. The paper contributes to the understanding of using deep learning for suicide risk detection in social media, showcasing the potential of combining CNN and LSTM architectures for improved results in this sensitive domain.

In their paper, titled “Identifying suicide attempts, ideation, and non-ideation in major depressive disorder from structural MRI data using deep learning” [25], the authors aimed to identify suicide risks in major depressive disorder (MDD) patients by analyzing structural MRI (sMRI) data through deep learning techniques. The dataset included 288 MDD patients, categorized into three groups: suicide attempts (SA), suicidal ideation (SI), and non-suicidal ideation/attempts (NS). Deep neural network models were developed for three classification tasks: SA versus SI, SA versus NS, and SI versus NS. The models achieved the highest accuracy of over 70%. The statistical analysis revealed specific brain regions and features that contributed significantly to the classification, shedding light on potential structural differences associated with suicidal behaviors in MDD patients. The study’s limitations included a relatively small dataset and the absence of certain behavioral dimensions. Nevertheless, the research highlights the potential of interpretable deep learning methods in identifying different subtypes of suicidal behaviors among MDD patients using sMRI data.

The paper “Prediction of suicidal ideation in children and adolescents using machine learning and deep learning algorithm: A case study in South Korea where suicide is the leading cause of death” [26] focuses on predicting suicidal ideation in children and adolescents using machine learning techniques. The study employs a dataset collected by the Korea Youth Policy Institute between 2017 and 2020 and utilizes various machine learning algorithms, including logistic regression, Random Forest, XGBoost, MLP, and CNN, to analyze the data. The results indicate that the CNN algorithm achieved the highest

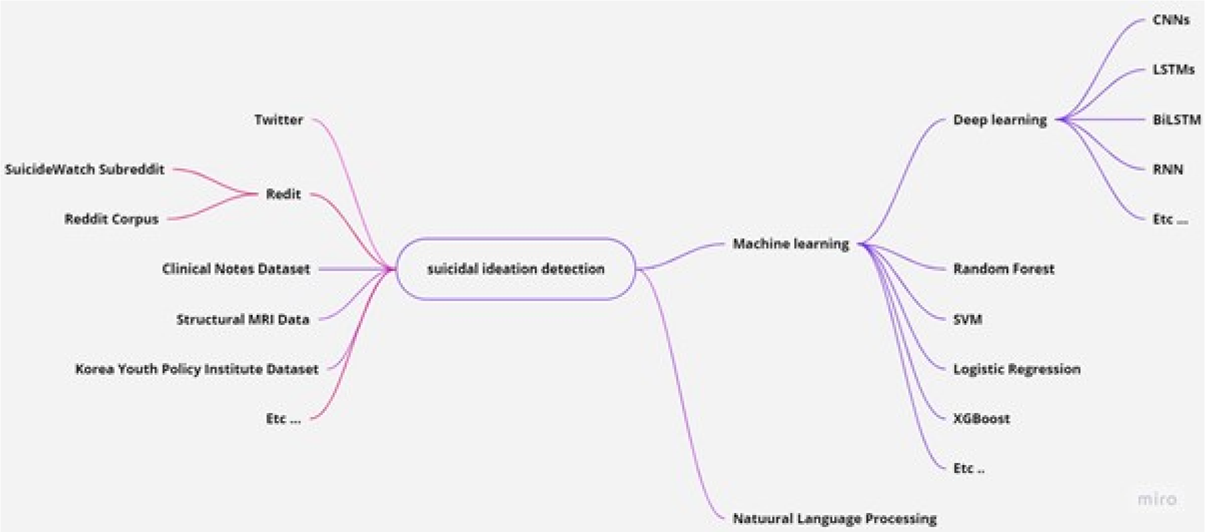


**Figure 4:** Categorization of Suicide Ideation Detection: Algorithms and Datasets. The right section illustrates algorithms and techniques, while the left section showcases the datasets utilized in the mentionned papaers.

prediction performance, with an 88.3% accuracy and 90.1% precision. The study also identifies key factors associated with increased suicidal ideation, such as sadness, depression, anxiety, loneliness, and name-calling. These findings emphasize the potential of machine learning in early identification and prevention of suicidal ideation in young individuals, highlighting the importance of timely intervention in suicide prevention efforts.

## 2.2. Advanced Algorithms: Combining NLP with Advanced Learning Methods

In the domain of suicide ideation detection, the integration of advanced algorithms has proven to be a game-changer. This sub-section delves into the synergy of Natural Language Processing with Machine techniques. By combining these powerful technologies, researchers have unlocked new horizons in the early identification and prevention of suicidal ideation across diverse data sources, including social media content and clinical notes. In this sub-section, we explore the innovative methods and approaches that leverage NLP's linguistic analysis capabilities, machine learning's predictive strength, and deep learning's nuanced understanding of complex textual data. Together, these elements empower more effective suicide risk assessment and intervention strategies, as elucidated in Recent advancements in suicide ideation detection have harnessed the power of Natural Language Processing and Machine Learning. These approaches have shown great promise: BiLSTM models achieved 93.6% accuracy, excelling with lengthy tweets; NLP combined with DL demonstrated an 84.15% accuracy for mental health risk detection; TCNN-MF-LA outperformed prior methods in Chinese social media; hybrid DL models with NLP preprocessing achieved high accuracy and F1 scores; XGBoost excelled at 96.33% accuracy with NLP features; multi-modal NLP features showed promise; and ensemble classifiers maintained consistency despite changing language and events. This interdisciplinary approach holds significant potential for advancing suicide ideation detection and early intervention efforts. Table 3. In a study by Aladag et al [27], the authors developed a technique to identify suicidal ideation in online forum posts. They gathered data from Redit including subreddits like "SuicideWatch," "ShowerThoughts," "Depression," and "Anxiety" on Reddit. After preparing the text data they employed natural language processing (NLP) methods along with machine learning algorithms such, as regression, random forest and support vector machines (SVM) to classify posts as either indicating tendencies or not based on linguistic and contextual attributes. The models demonstrated F1 scores and accuracy with regression and SVM achieving an F1 score of 92. The authors also acknowledged the limitations of their study, including language restrictions. Suggested research avenues to enhance their approach. In a study by Aladag et al [27], the authors developed a technique to identify suicidal ideation in online forum posts. They gathered data from Redit including subreddits like "SuicideWatch," "ShowerThoughts," "Depression," and "Anxiety" on Reddit. After preparing the text data they employed natural language processing (NLP) methods along with machine learning algorithms such, as regression, random forest and support vector machines (SVM) to classify posts as either indicating tendencies or not based on linguistic and contextual attributes. The models demonstrated F1 scores and accuracy with regression and SVM achieving an F1 score of 92. The authors also acknowledged the limitations of their study, including language restrictions. Suggested research avenues to enhance their approach. Paper [28] presents a comparative analysis of methods for detecting suicidal ideation in user tweets using Natural Language Processing, Machine Learning, and Deep Learning (DL) techniques. The study utilizes a dataset of 49,178 tweets, created by collecting live tweets containing suicide-indicative and non-suicidal keywords. Five ML algorithms and four DL models are trained on this dataset, with the best-performing model being BiLSTM, achieving an accuracy of 93.6% an F1 score of 0.93. The paper highlights the advantages of BiLSTM in handling lengthy tweets effectively due to its ability to capture forward-backward dependencies. Additionally, it mentions the potential for further improvement using different feature extraction techniques and the need for multi-class datasets in future research efforts in this critical area of early suicide detection. In their paper titled "A computational model for assisting individuals with suicidal ideation based on context histories" [29], authors leverage deep learning algorithms, particularly neural networks with multiple layers, in conjunction with Natural Language Processing techniques for the detection of suicidal ideation and sentiment prediction within social media content, with a specific focus on Twitter. These deep learning methods, combined with NLP, facilitate the automatic extraction of intricate patterns from vast datasets of user-generated content. Importantly, the statistical results reveal promising outcomes in identifying the "Risk" cases using simulated data, achieving an average accuracy of 84.15%, along with 84.15% for recall and 91.45% for F1-Score metrics. In [30] Li et al employed Natural Language Processing techniques to develop a deep learning model for suicide detection on Chinese social media platforms. They utilize NLP methods to preprocess and analyze text data, extracting multiple text-based features. The model employs convolutional neural networks and other deep learning algorithms for text classification. The study presents statistically significant results, demonstrating that their proposed TCNN-MF-LA model, which incorporates NLP-driven features and label association mechanisms, outperforms previous models in accurately detecting suicidal ideation. This research is based on a constructed Chinese social media suicide detection dataset, showcasing the authors' reliance on NLP for understanding and addressing mental health concerns in online text data. In the meanwhile, authors in [31]employed a hybrid deep learning model to address the classification and prediction of suicidal ideation from social network data. The methodology involves extensive Natural Language Processing (NLP) techniques, including text preprocessing tasks such as lemmatization, symbol and stopwords removal, and tokenization. Two word embedding techniques, Glove and Random, are applied to convert textual data into numerical vectors, a crucial step in NLP. Various deep learning algorithms, such as Convolutional Neural Networks and Long Short-Term Memory, both individually and in combinations, are utilized. The study evaluates the model's performance using statistical metrics like accuracy, precision, recall, F1 score, and specificity. The dataset comprises 20,000 posts from the SuicideWatch subreddit, with 10,000 labeled as suicidal and 10,000 as non-suicidal. The results indicate that the proposed model, particularly when tuned with Glove embeddings, achieves high accuracy, precision, and F1 scores. NLP plays a pivotal role in preprocessing the textual data, enabling effective classification and prediction of suicidal ideation from social network posts. The paper "Detecting suicidality on social media: Machine learning at rescue"[32], employs Natural Language Processing techniques as a fundamental part of its methodology for detecting suicidality on social media. It utilizes various textual features, including TFIDF (Term Frequency-Inverse Document Frequency), latent semantic indexing features, and average risk similarity features, all of which are common in NLP. These NLP techniques help analyze and extract meaningful information from the text data found in social media posts. The study applies machine learning algorithms, such as Support Vector Machine, Random Forest, and Extreme Gradient Boosting, to train a suicide prediction model based on these NLP-derived features. The statistical results indicate that the XGBoost-based model outperforms the others, achieving an accuracy of 96.33%. The dataset used likely comprises social media posts with labeled suicidal risk classes, enabling the algorithms to learn patterns in the text data associated with different risk levels. In study[33] natural language processing techniques were employed to extract linguistic characteristics, sentiment analysis, and emotional analysis (EA) from textual data. These NLP-driven analyses contributed to the development of multi-modal features for suicide ideation detection. Several machine learning algorithms were applied to these features, resulting in the following statistical results: The LDA with the LR classification model achieved an F1-score of 0.85 and an accuracy of 86 percent (86 percent, 0.85), indicating high performance. Trigram + TF-IDF with RF classifier achieved the second-highest accuracy at 77 percent. Temporal analysis as a single feature using SVM achieved an accuracy of 74 percent. Sentiment Analysis, when utilized with the XGBoost model, reached a 75 percent accuracy (75, 0.74). Emotional Analysis (EA), which assesses users' emotional states based on emoticon usage, contributed significantly. Although EA alone achieved an accuracy of 72 percent (72 percent, 0.71), it performed exceptionally well when combined with other classifiers, resulting in an 87 percent accuracy (87, 0.87). In [34] Burnap et al presented a method for classifying suicide-related communication in social media using machine learning algorithms and natural language processing techniques. The study conducted a 12-month analysis, splitting it into two tasks: binary classification (identifying whether text is suicidal) and a more detailed 7-class classification. For the binary task, 85% accuracy was achieved in confirming suicidal content by human annotators on a sample of 2000 tweets classified by their ensemble method. In the 7-class task, 65.29% accuracy was achieved for classifying suicidal ideation. The dataset consisted of social media posts, and the ensemble classifier demonstrated consistency over the study period despite various events and language changes. In their study titled "Bootstrapping semi-supervised annotation method for potential suicidal messages", authors in [35] focused on evaluating the performance of their semi-supervised learning system in detecting messages of depression or suicidal ideation. The main statistical result they emphasized was the macro F1 score, which ranged from 0.78 to 0.81. These F1 scores were found to be promising, as they were close to the mutual agreement reached by human reviewers, as measured by Cohen's Kappa of 0.86. This strong performance was achieved when the Life Corpus was expanded by adding 171 samples from the Reddit Corpus, where annotators had reached mutual agreement. These results validate the effectiveness of the chosen semi-supervised Bootstrapping Uncertainty Sampling methodology in improving the automatic system's ability to detect such messages. Recent advancements in suicide ideation detection have harnessed the power of Natural Language Processing and Machine Learning. These approaches have shown great promise: BiLSTM models achieved 93.6% accuracy, excelling with lengthy tweets; NLP combined with DL demonstrated an 84.15% accuracy for mental health risk detection; TCNN-MF-LA outperformed prior methods in Chinese social media; hybrid DL models with NLP preprocessing achieved high accuracy and F1 scores; XGBoost excelled at 96.33% accuracy with NLP features; multi-modal NLP features showed promise; and ensemble classifiers maintained consistency despite changing language and events. This interdisciplinary approach holds significant potential for advancing suicide ideation detection and early intervention efforts.

## 2.3. Common features of the suicidal patients

In this paper, we will introduce the concept of "suicidal patients" for the purpose of identifing those with suicidal ideations, encompassing individuals, users, patients, or participants within social networks. In the subsequent section, we will highlight the main characteristics employed in the identification of these individuals with suicidal tendencies, which will be summirized in Table 4 bellow

The summarized results in the table showcase a multifaceted approach to identify and classify individuals with suicidal tendencies. Common features for classification encompass linguistic analysis, semantic assessment, machine learning algorithms, and text-based indicators. Some papers stress the importance of specific keywords and sentiment analysis, while others employ deep learning models or structural MRI data. These findings underscore the complexity of detecting individuals at risk of suicidal thoughts or behavior and highlight the need for comprehensive methods across various contexts.

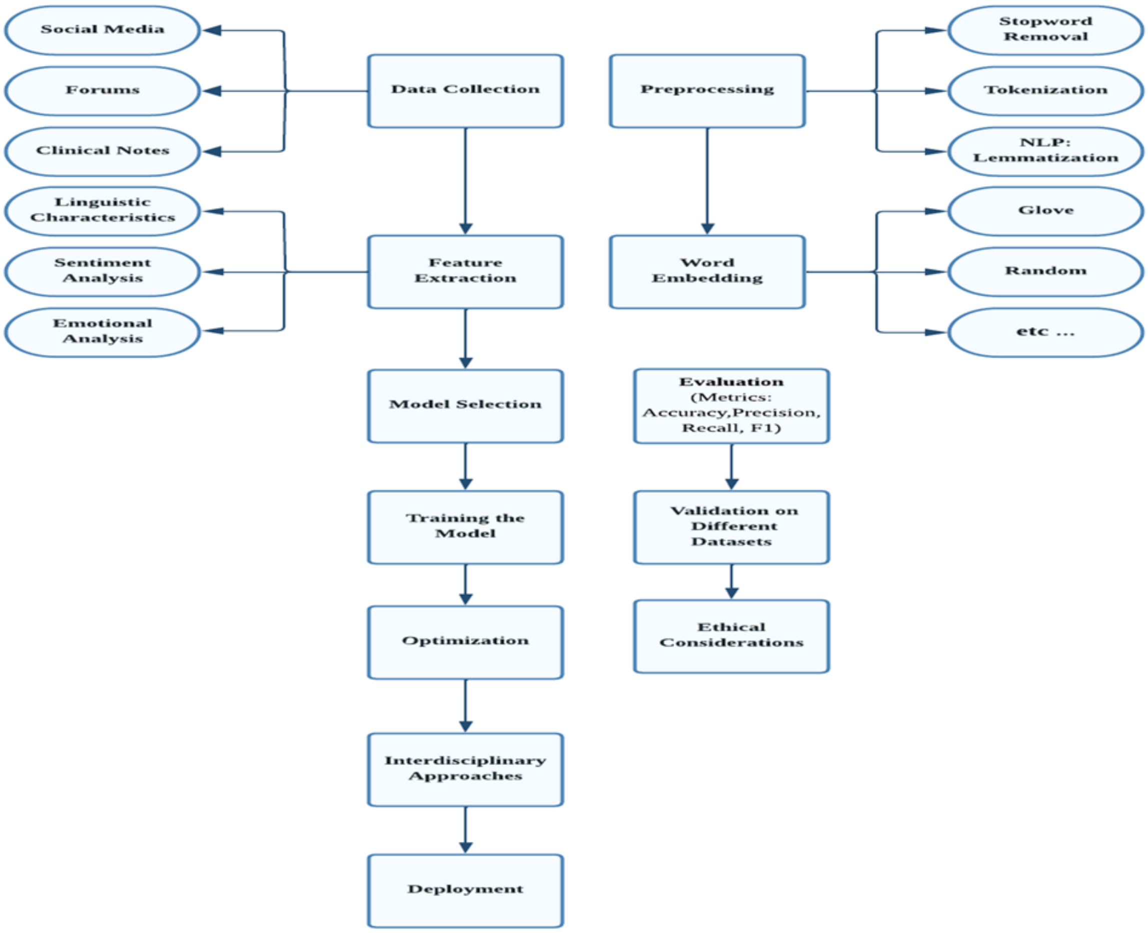


**Figure 5:** Machine Learning main architecture used for suicidal ideation detection.

## 3. Results and discussion

The reviewed studies collectively demonstrate the efficacy of various Natural Language Processing, Machine Learning, and Deep Learning approaches in suicide ideation detection. NLP-driven features and label association mechanisms proved effective, with Li et al [30].'s TCNN-MF-LA model outperforming prior models in Chinese social media. Hybrid DL models, as showcased by Chadha and Kaushik, integrating NLP techniques, achieved high accuracy and F1 scores in social network data. Rabani et al.'s study, combining NLP-derived features with ML algorithms, yielded impressive accuracy with XGBoost. Multi-modal NLP features, as shown by Chatterjee et al., demonstrated high accuracy and F1 scores, emphasizing their effectiveness. DL methods, such as Renjith et al.'s BiLSTM model, proved adept at handling lengthy text data, achieving a 93.6% accuracy. Rentz et al.'s neural network-based model showcased DL's potential for clinical applications with an 84.15% accuracy in identifying "Risk" cases. Tadesse et al. highlighted DL's power in detecting suicide ideation in social media forums. Hu et al. extended DL to structural MRI data for identifying suicide attempts and ideation. Shin and Kim employed ML and DL algorithms, with CNN achieving 88.3% accuracy for predicting suicidal ideation in children and adolescents. Chatterjee et al.'s multi-modal approach and Burnap et al.'s ensemble methods demonstrated robustness, while Acuña Caicedo et al.'s semi-supervised learning system showed promise. Collectively, these studies underscore the potential of these interdisciplinary approaches for early intervention and improved mental health outcomes.

**Table 3**
Summary of Deep Learning-Based Approaches and Performance Metrics in Suicidal Ideation Prediction

| Study | Approach | Dataset Size | Best Model | Best Model Accuracy | Key Findings |
|---|---|---|---|---|---|
| Aladag et al [27] | Logistic Regression, RF, SVM | Forum posts (thousands of them) | Logistic regression, SVM | F1 score of 92% | provide immediate support to individuals at risk of suicide |
| Haque et al [28] | NLP, ML, DL | 49,178 tweets | BiLSTM | 93.6% | Effective for lengthy tweets |
| Rentz et al [29] | DL, NLP | Simulated data | Neural Networks | 84.15% | Promising results for "Risk" case detection |
| Li et al [30] | NLP, DL | Chinese social media data | TCNN-MF-LA | 85.50% | Effective for Chinese social media |
| Chadha et al [31] | NLP, DL | 20,000 Reddit posts | Hybrid DL models | 88.48% | Effective for social network data |
| Rabani et al [32] | NLP, ML | Social media posts | XGBoost | 96.33% | High accuracy with NLP features |
| Chatterjee et al [33] | NLP, ML | Textual data | LDA, LR | F1-score of 0.85, 86% accuracy | Effectiveness of multi-modal NLP features |
| Burnap et al [34] | NLP, ML | Social media posts | Ensemble method | 85% binary classification, 65.29% 7-class classification | Consistency despite changing language and events |
| Acuña et al [35] | Semi-supervised Learning, NLP | Life Corpus + Reddit Corpus | SVM classifier, semi-supervised method | 0.80 of macro f1 score | Semi-supervised system achieved promising macro F1 scores, close to human reviewer agreement. |

| **Common Features for Classification** | **Mentioned Papers** |
|---|---|
| Linguistic Elements | [16], [19], [20], [22], [27] |
| Machine Learning Algorithms | [17], [18], [20], [7], [22], [27] |
| Semantic Analysis | [17] |
| Specific Vocabulary and Keywords | [18] |
| Sentiment Analysis | [19], [22] |
| Social Media Interactions | [19], [20], [22] |
| Text-Based Features | [22] |
| LIWC Analysis | [22], [27] |
| Term Frequency-Inverse Document Frequency | [27] |
| SMOTE (Synthetic Minority Over-sampling Technique) | [27] |

**Table 4**
Common Features Employed for the Classification of Suicidal Tendencies Across Studies.

## 4. Conclusion

In conclusion, the extensive review of studies in this literature highlights the remarkable progress made in the field of suicide ideation detection using a variety of advanced algorithms and techniques. Natural Language Processing, Machine Learning, and Deep Learning methods have proven to be invaluable tools in identifying and understanding suicidal ideation across diverse data sources, including social media content and clinical notes. These studies collectively emphasize the potential of interdisciplinary approaches, combining NLP's linguistic analysis capabilities, ML's predictive strength, and DL's nuanced understanding of complex textual data. The results showcase the high accuracy achieved by various models, indicating their effectiveness in identifying early signs of suicidal thoughts. These advancements hold significant promise for enhancing suicide prevention efforts and early intervention strategies, ultimately contributing to improved mental health outcomes for individuals at risk. However, it is essential to acknowledge the need for further research, larger datasets, and ongoing refinement of these algorithms to ensure their continued effectiveness and real-world applicability in suicide prevention and mental health support.

## References

[1] The world health organization (WHO), "Suicide." Accessed: Aug. 28, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[2] F. R. Yu and Y. He, "Introduction to Machine Learning," 2019, pp. 1–13. doi: 10.1007/978-3-030-10546-4_1.

[3] N. V. Nagirimadugu and S. Tippireddy, "Recommendations for Integrating the Fundamentals of Machine Learning Into Medical Curricula," Academic Medicine, vol. 96, no. 9, pp. 1230–1230, Sep. 2021, doi: 10.1097/ACM.0000000000004192.

[4] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," APSIPA Trans Signal Inf Process, vol. 3, no. 1, 2014, doi: 10.1017/atsip.2013.9.

[5] B. Wen et al., "Deep Learning in Proteomics," Proteomics, vol. 20, no. 21–22, Nov. 2020, doi: 10.1002/pmic.201900335.

[6] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," Comput Intell Neurosci, vol. 2018, pp. 1–13, 2018, doi: 10.1155/2018/7068349.

[7] C. Cao et al., "Deep Learning and Its Applications in Biomedicine," Genomics Proteomics Bioinformatics, vol. 16, no. 1, pp. 17–32, Feb. 2018, doi: 10.1016/j.gpb.2017.07.003.

[8] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," Med Image Anal, vol. 65, p. 101759, Oct. 2020, doi: 10.1016/j.media.2020.101759.

[9] R. Dias and A. Torkamani, "Artificial intelligence in clinical and genomic diagnostics," Genome Med, vol. 11, no. 1, p. 70, Dec. 2019, doi: 10.1186/s13073-019-0689-8.

[10] Z. Taskin and U. Al, "Natural language processing applications in library and information science," Online Information Review, vol. 43, no. 4, pp. 676–690, Aug. 2019, doi: 10.1108/OIR-07-2018-0217.

[11] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "The Woman Worked as a Babysitter: On Biases in Language Generation," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3405–3410. doi: 10.18653/v1/D19-1339.

[12] Y. Ma, "A Study of Ethical Issues in Natural Language Processing with Artificial Intelligence," Journal of Computer Science and Technology Studies, vol. 5, no. 1, pp. 52–56, Mar. 2023, doi: 10.32996/jcsts.2023.5.1.7.

[13] L. Weidinger et al., "Ethical and social risks of harm from Language Models," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.04359

[14] K. P. Anuradha, An Introduction Natural Language Processing, vol. 1, no. 1. 2019. [Online]. Available: http://restpublisher.com/book-

[15] J. Lee and T. Y. Pak, "Machine learning prediction of suicidal ideation, planning, and attempt among Korean adults: A population-based study," SSM Popul Health, vol. 19, Sep. 2022, doi: 10.1016/j.ssmph.2022.101231.

[16] R. W. Acuña Caicedo, J. M. Gómez Soriano, and H. A. Melgar Sasieta, "Assessment of supervised

classifiers for the task of detecting messages with suicidal ideation," Heliyon, vol. 6, no. 8, Aug. 2020, doi: 10.1016/j.heliyon.2020.e04412.

[17] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks," in Procedia Computer Science, Elsevier B.V., 2017, pp. 65–72. doi: 10.1016/j.procs.2017.08.290.

[18] A. C. De Oliveira, E. J. S. Diniz, S. Teixeira, and A. S. Teles, "How can machine learning identify suicidal ideation from user's texts? Towards the explanation of the Boamente system," in Procedia Computer Science, Elsevier B.V., 2022, pp. 141–150. doi: 10.1016/j.procs.2022.09.093.

[19] S. Renjith, A. Abraham, S. B. Jyothi, L. Chandran, and J. Thomson, "An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, pp. 9564–9575, Nov. 2022, doi: 10.1016/j.jksuci.2021.11.010.

[20] B. Priyamvada et al., "Stacked CNN - LSTM approach for prediction of suicidal ideation on social media," Multimed Tools Appl, Jul. 2023, doi: 10.1007/s11042-023-14431-z.

[21] M. Cusick et al., "Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation," J Psychiatr Res, vol. 136, pp. 95–102, Apr. 2021, doi: 10.1016/j.jpsychires.2021.01.052.

[22] T. H. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, and Z. A. T. Ahmed, "Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models," Int J Environ Res Public Health, vol. 19, no. 19, Oct. 2022, doi: 10.3390/ijerph191912635.

[23] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using deep learning," Algorithms, vol. 13, no. 1, Jan. 2020, doi: 10.3390/a13010007.

[24] A. C. Gyllensten and M. Sahlgren, "Measuring Issue Ownership using Word Embeddings," in WASSA 2018 - 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Proceedings of the Workshop, Association for Computational Linguistics (ACL), 2018, pp. 149–155. doi: 10.18653/v1/P17.

[25] J. Hu et al., "Identifying suicide attempts, ideation, and non-ideation in major depressive disorder from structural MRI data using deep learning," Asian J Psychiatr, vol. 82, Apr. 2023, doi: 10.1016/j.ajp.2023.103511.

[26] S. Shin and K. Kim, "Prediction of suicidal ideation in children and adolescents using machine learning and deep learning algorithm: A case study in South Korea where suicide is the leading cause of death," Asian J Psychiatr, vol. 88, Oct. 2023, doi: 10.1016/j.ajp.2023.103725.

[27] A. E. Aladag, S. Muderrisoglu, N. B. Akbas, O. Zahmacioglu, and H. O. Bingol, "Detecting suicidal ideation on forums: Proof-of-concept study," J Med Internet Res, vol. 20, no. 6, Jun. 2018, doi: 10.2196/jmir.9840.

[28] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning," Technologies (Basel), vol. 10, no. 3, Jun. 2022, doi: 10.3390/technologies10030057.

[29] D. M. Rentz, W. F. Heckler, and J. L. V. Barbosa, "A computational model for assisting individuals with suicidal ideation based on context histories," Univers Access Inf Soc, 2023, doi: 10.1007/s10209-023-00991-2.

[30] Z. Li, W. Cheng, J. Zhou, Z. An, and B. Hu, "Deep learning model with multi-feature fusion and label association for suicide detection," Multimed Syst, Aug. 2023, doi: 10.1007/s00530-023-01090-1.

[31] A. Chadha and B. Kaushik, "A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data," New Gener Comput, vol. 40, no. 4, pp. 889–914, Dec. 2022, doi: 10.1007/s00354-022-00191-1.

[32] S. T. Rabani, A. M. Ud Din Khanday, Q. R. Khan, U. A. Hajam, A. S. Imran, and Z. Kastrati, "Detecting suicidality on social media: Machine learning at rescue," Egyptian Informatics Journal, vol. 24, no. 2, pp. 291–302, Jul. 2023, doi: 10.1016/j.eij.2023.04.003.

[33] M. Chatterjee, P. Kumar, P. Samanta, and D. Sarkar, "Suicide ideation detection from online social media: A multi-modal feature based technique," International Journal of Information Management Data Insights, vol. 2, no. 2, Nov. 2022, doi: 10.1016/j.jjimei.2022.100103.

[34] P. Burnap, G. Colombo, R. Amery, A. Hodorog, and J. Scourfield, "Multi-class machine classification of suicide-related communication on Twitter," Online Soc Netw Media, vol. 2, pp. 32–44, Aug. 2017, doi: 10.1016/j.osnem.2017.08.001.

[35] R. W. Acuña Caicedo, J. M. Gómez Soriano, and H. A. Melgar Sasieta, "Bootstrapping semi-supervised annotation method for potential suicidal messages," Internet Interventions, vol. 28. Elsevier B.V., Apr. 01, 2022. doi: 10.1016/j.invent.2022.100519.

# AI-Driven Cybersecurity Orchestration and Automation

Touil Lokman[1], Betouil Ali Abdletif[1,2]

[1] *Computer Science Department , University Chadli Bendjedid, El Tarf , Algeria*
[2] *Laboratoire d'informatique et des mathématiques appliquées (LIMA), El Tarf, Algeria*

## Abstract

The following paper researches how Artificial Intelligence (AI) is being integrated into cybersecurity, with a special focus on SOAR (Security Orchestration, Automation, and Response) as a strategy to counter sophisticated, ever-evolving cyber threats. It introduces AI-driven SOAR platforms that enhance threat detection, incident response, and overall security posture through adaptive and autonomous systems. Key topics discussed include the design of frameworks for incorporating AI into Security Information and Event Management (SIEM) systems, challenges in playbook automation, and ethical concerns related to deploying AI in cybersecurity. The paper also examines AI applications in decision-making processes, predictive modeling, and proactive security measures, addressing limitations such as integration complexity and privacy concerns. It concludes by providing insights into the future trajectory of AI in cybersecurity, emphasizing its transformational potential and the need to balance human expertise with AI-driven automation

## Keywords

Artificial Intelligence, cybersecurity, SOAR, SIEM, threat detection, incident response, automation, predictive modeling, privacy concerns, security posture, adaptive systems.

## 1. Introduction

Orchestration and automation technologies contribute to the realization of a faster threat detection and incident response against cybersecurity attacks which is crucial in an era of advanced sophisticated security threats. Leveraging the advancement of Artificial Intelligence (AI), AI-driven cybersecurity orchestration and automation can support robust orchestration and automation in a self-healing, intelligent way.

The role of Artificial Intelligence (AI) technology in cybersecurity is attracting increasing interest. The increasing sophistication, volume, and prevalence of cyberattacks have generated a relentless demand for advanced detection, monitoring, and mitigation solutions. These AI-driven solutions can add an extra layer of defence by monitoring connected IT assets and detecting any anomalies resulting from attacks, thus improving threat detection [2]. In response to the need for new AI capabilities in cybersecurity orchestration and automation, well-known limitations and challenges of AI for cybersecurity orchestration and automation are published. These limitations act as key drivers for further research in the development of the architecture, processes, and mechanisms of new AI-driven, intelligent cybersecurity orchestration, and automation solutions, as well as the integration of these solutions into SIEM technologies and IDE technology stacks.[1]

### 1.1. Background and Significance :

Cybersecurity ensures the confidentiality, integrity, and availability of digital data and related services. This procedure has transformed in recent years as a result of major improvements in digital infrastructure and the global acceptance of advanced technologies. The cybersecurity realm is now confronted with increasingly complex cyber threats, new attack patterns, and sophisticated technologies [1]. As cyber breaches proliferate, the conventional manual Cybersecurity Orchestration and Automation (COrA) systems would prove to be unsustainable in ensuring complete cyber protection, given the rapidly growing threat surfaces. The existing solutions are either ad hoc, Semi-Automated, or zero/partial awareness automated approaches operating in a reactive manner without any integration of Artificial Intelligence technologies. A complete end-to-end threat response solution is proposed that incorporates all the major aspects of the cybersecurity response ecosystem. This orchestration approach aims to tackle the obstacles of the current state-of-the-art threat response techniques integrating awareness and autonomous decision-making capabilities [2].

The proposed AI-Driven COrA that supports the continuous detection, investigation, containment, eradication, and recovery from cyber threats. These activities utilize a closed-loop procedural architecture for adaptive cyclic orchestration. This procedure radically distinguishes between the System Deployed Agent and an AI-Controlled Orchestrator agent that maintains an abstracted architecture of the threat response landscape. The AI-Controlled Orchestrator Agent is responsible for managing cyber operations, resource allocation, and the selection of automated response tactics, techniques, and procedures (TTPs). These operations blend human intelligence, machine learning, and advanced agent technologies within a modeling framework for representational awareness of the threat response landscape

### 1.2. Research Objectives :

The expansive adoption of software applications and internet-based services has engendered the need to safeguard trillions of dollars in assets from cyber threats. Such threats can be materialized either through external hackers, disgruntled employees, or apparent access types. The cybersecurity orchestration and automation paradigm provides faster, easier, and more expansive security capabilities, especially under the remedy of alerts initiated from the cybersecurity knowledge base.

The process of analyzing and remediating alerts through the investigation of log files from multiple security devices, correlating the alerts in real-time with security events, and implementing remediation activities as a response to alerts is considerably complicated and labor-intensive. Thus, it calls for effective and automatic solutions to enable time-efficient forensic analysis and recovery, targeting the business development of small security firms using orchestration/automation platforms.

A framework concept and architecture targeting the orchestration and automation of cybersecurity-related applications for the automated solution of alerts originated or linked with a SIEM security platform is proposed. The framework offers a comprehensive view of the security applications at the enterprise level while automating remediation of alerts. The objectives include integration with different SIEM solutions and regulations, understanding and detecting malicious actions, including the vast knowledge from online sources even regarding security vulnerability in applications or devices, assessing risk and cost/benefit analysis taking into consideration each entity in the enterprise framework, and offering remediation proposals.

Furthermore, the feasibility of proposed actions is sought to be assessed, starting engagement with remediation actions only if feasible, combining different aspects in accordance with enterprise policy regarding risk acceptance according to internal/external business competition in correspondence to asset protection.

## 2. Fundamentals of AI in Cybersecurity :

### 2.1. Overview of AI in Cybersecurity:

The breadth of applications of AI in the cybersecurity domain is wide and varied in terms of types, processes, and outcomes, and is classified into three groups: actionable outcomes, broad application areas, and AI technologies as enablers of outcomes [1]. These broad categories provide insight into the applicability and implications of AI technologies in cybersecurity. To enhance understanding of the implications of a specific AI technology, a more fine-grained inquiry is conducted by identifying and analyzing the technology processes and outcomes. These processes and outcomes help to understand the intersection of technology and social contexts, such as the nature, impact, and consequences of the application of technology [3]. The exploratory analysis is based on a systematic review of the literature on AI applications in cybersecurity from the years 2010 to 2021. AI technologies applied at multiple stages of application life cycles/dimensions and the security types are presented.

To foster further discussions and inquiries on the applicability, impact, and implications of AI in cybersecurity, a theoretical framework is constructed by synthesizing the findings from three lenses of inquiry comprising AI technologies, security types, and technology outcomes. This framework is used to analyze and discuss emerging knowledge in regards to patterns, dependencies, and relationships. AI is rapidly being integrated into various industry sectors for cybersecurity applications to combat the increasing threat of cybersecurity attacks. The pervasiveness of connectivity, use of portable devices, deployment of various technologies, digitalization, use of cloud, artificial intelligence-driven technologies, Internet of Things, big data, BYOD policies have increased vulnerability of networks and systems to cybersecurity threats and attacks.

### 2.2. Aplications of AI in Security Operations:

A telemetry and events correlation step precedes any incident response action. This action is commonly achieved by a (SIEM) solution with defined rules and conditions of detection. However, log analysis and event detection usually require every component's own context, potentially missing important detections. Therefore, organizations are investing in Threat Intelligence platforms (TIP), which gather threat-related data from multiple sources, normalizing it and providing it with context and scoring rules [2].

## 3. Security Orchestration , Automation and Response (SOAR):

Security Orchestration, Automation, and Response (SOAR) is an emerging concept to assist cybersecurity experts and professionals in automatically responding to threats, handling alerts, and profiting from additionally required information [4]. Security orchestration incorporates security information and event management (SIEM) systems and threat intelligence solutions, wherein siloed security solutions share information and threat intelligence. One of the SOAR platforms is the Microsoft Sentinel Security Orchestration, Automation, and Response (SOAR) solution, which automates alerts and tasks from Microsoft Sentinel, Microsoft Defender for Cloud, and Microsoft Defender for Endpoint.

Some of the key benefits of implementing SOAR solutions in a security operation center (SOC) organization are to accelerate the investigation and response processes and improve the security posture and efficiency of security analysts by reducing false positives and noise [5]. The concept of security orchestration and orchestration platforms is gaining momentum in the cybersecurity domain. New vendors are launching security orchestration and automation platforms; however, there is a degree of confusion regarding the term itself, possibly due to the inception of new buzzwords and variations of security orchestration. A formal definition and explanation of crucial terms and concepts are provided to better comprehend security orchestration.

### 3.1. Definition of Components of SOAR :

SOAR refers to Security Orchestration, Automation, and Response. SOAR platforms enable security teams to respond to incidents quickly and optimally, assuring time-determined financial impact mitigation [5]. SOAR platforms enhance the abilities of security solutions by enriching the gathering of alerts and situational awareness. SOAR is often seen as a shift towards the automation of security analyses, administration, and remediation tasks. SOAR should be viewed as a platform that encompasses and makes use of various solutions (e.g., security and observability tools) in performing security operations. SOAR platforms encompass the following core components: security orchestration, event management, security automation, playbooks, incident management, runbooks, threat intelligence management, and case management. A detailed description of these core components follows.

Security orchestration refers to the integration of security and information technology tools [or domain-specific devices] designed to facilitate, streamline, and centrally manage security automation and processes [4]. Event management refers to the management of event collection, categorization, prioritizing, and resolution of security alerts from various information systems (e.g., network devices, firewalls, security gateways, servers, databases, applications) in real or near-real time. Security automation refers to the use of information technology in place of manual processes for the purpose of security event management, vulnerability management, cyber incident response, and other security activities.

### 3.2. Benefits of SOAR Platforms :

Security Orchestration, Automation, and Response (SOAR) platforms are gaining traction in security operations centers (SOCs) of all sizes for integrating disparate security tools, automating repetitive manual tasks, and allowing security analysts to prioritize the handling of alerts by threat level. Within a managed security services provider (MSSP) environment, SOAR platforms also offer the ability to easily coordinate the sharing of security data between the MSSP and customers, as well as among customers of the MSSP, creating a richer, more contextual perspective of security incidents that enhances the efficiency and efficacy of a security team [4]. They provide value through the following benefits:

1. Centralized Security Incident Management SOAR platforms enable security teams to significantly reduce the collection and collation time of data relevant to investigating security incidents by allowing it to come from multiple, disparate sources through an easy-to-use interface and present the information in a concise, pre-defined format. By automating the creation of an incident, a team can save substantial time and effort previously spent on numerous mundane, repetitive tasks in investigating incidents. These process enhancements are magnified when aggregated across a security team as a whole.

2. Focus on the Most Critical Security Incidents SOAR platforms create severity assessments backed by contextual information that inform security teams on how to prioritize and respond to incidents. By setting clear thresholds on this information, security alerts can be categorized by threat level and escalated appropriately. Security incidents are more likely to make it into the hands of the analysts who are capable of handling them, resulting in a net gain in productivity for the security team. Additionally, the maturity of incident categorization across a team can inform the creation of analytics to proactively measure performance.

3. Centralized Reporting and Metric Generation SOAR platforms drive standardization in how incidents are captured and responded to, which enables security teams to centrally aggregate this information and generate reports to offer insight into repeated, high-fidelity security alerts, team performance, and process efficiency [5].

## 4. AI-Driven SOAR Platforms :

AI-Driven SOAR Platforms are Security Orchestration, Automation, and Response (SOAR) Platforms that incorporate artificial intelligence capabilities to enhance data analysis, threat detection, incident response, or other cybersecurity functions. Organizations invest in multiple different security tools, each focused on a specific capability such as network or endpoint detection, investigation, protection, and remediation. While there can be redundancy among these tools in terms of coverage, it is important to leverage each tool's unique capabilities in a comprehensive and effective security program. Cybersecurity professionals in Operations or Security Operation Centers (SOCs) are responsible for enabling the integration among individual security tools through processes and procedures called "playbooks." Playbooks describe the actions taken to investigate, respond to, or remediate issues using different security tools . This integration is often facilitated by Security Orchestration, Automation, and Response (SOAR) platforms.

SOAR platforms allow for the automation of playbook actions to improve the consistency and efficiency of investigations or responses with a goal of reducing exposure time and expediting remediation. While playbooks can be relatively simplistic, leveraging a single tool to perform a straightforward action, current playbook design and development workflows are often overly dependent on cybersecurity professionals' expertise and experience with tools and threats. The burden is typically placed solely on these operators to take time away from investigating, monitoring, or responding to potential incidents in order to develop, tune, test, and maintain playbooks in a SOAR platform while ensuring the integrity of these functions is not negatively affected by the automation [6]. Despite continued advancements in threat intelligence, detection, and security tooling, the current threat landscape continues to grow in scale and sophistication presenting a significant burden on existing security operations. The broader adoption of artificial intelligence, machine learning, and automation in response to this increased burden could help alleviate the skills gap and improve overall efficiency.

### 4.1. Integration with Existing Security Tools :

To address the challenges of integrating multiple existing security tools and technologies in organization cybersecurity infrastructures, AI-Driven SOAR Platforms must provide out-of-the-box integration with various existing security tools. Such integration is facilitated through Application Programming Interfaces (APIs) that promote synchrony and interoperability [7].

Various Existing Security Technologies and Tools: In the cybersecurity infrastructure of organizations, diverse security technologies and tools are utilized for detecting, analyzing, prioritizing, and responding to security events. The effectiveness of each security tool is measured based on the synergies between multiple different tools. The tasks that cannot be performed by one tool can be executed by another tool. Various security technologies and tools are available or can be developed to cater to different organization needs. AI-Driven SOAR Platforms create synergy through aggregated integration and cohesive integration between the multiple existing security tools and technologies in organizations.

Out-of-the-Box Integration of Security Technologies and Tools: AI-Driven SOAR Platforms must provide seamless out-of-the-box integration with various existing security technologies and tools. Ideally, such integration is performed through pre-built connectors, which consist of defined APIs. Application Programming Interfaces (APIs) are identified methodologies and tools that enable technologies or tools to connect and communicate with one another through predefined protocols [5]. In the context of cybersecurity, APIs facilitate the establishment of systematic connectivity between security technologies and tools in cybersecurity infrastructure.

## 5. AI Algorithms in SOAR Systems :

Three types of algorithms normally run on the underlying Platform Processing Engine (PPE) of SOAR tools that utilize AI capabilities: AI ML algorithms, Intelligent Decision Support algorithms, and Deep Learning algorithms . AI ML algorithms allow for the automatic building of probabilistic or confidence estimates about objects, topics, or events by learning from historical examples in human-built databases referred to as "training sets." Thus as event allegations known to be "good" or "bad" for a system are reviewed (by humans) a corresponding example set of prior events is gradually built up. Subsequently, trained examples are used for building an operational AI algorithm (known as a "classifier") that estimates either the probability of such scenarios recurring or the confidence of those estimates. AI ML algorithms are now also extensively used within large email systems and proposed for use within other cyber domains like bot detection, "good" URL versus phishing URL detection, and spam detection [8]. Generally, when these detectors are up and running people using the system notice a dramatic decrease in the threat detection they see, however, a 2-to-10 times gain in detection efficiency is typical. On the other side of performance, if used excessively incorrectly such as with an overly aggressive threshold setting or after such a system was inappropriately tuned it could lead to a "hold onto your hat" situation for the recipients of alerts.

Intelligent Decision Support algorithms provide semi-automated human decision support scenarios. These algorithms have been widely used for the monitoring of diverse operational domains such as air defense systems, fault detection in aircraft, flight route selection, and personnel scheduling. Typically a human is provided with a graphical representation of the operational situation along with an automated inference of the state of that situation based on a set of human-built rules and historical outcomes. It is then up to the recipient of this output to decide how best to act on that situation. If a high degree of abstraction is used in the situation being presented it allows the human operator involved to concentrate on the trees without worrying about the forest. Such algorithms provide a dramatic enhancement in human decision performance and have been shown in multiple domains to both increase the correctness of decisions and speed them up.

Deep Learning is a new architectural approach to implementing AI capabilities in which one does not build databases that can be easily imported to a PPE, but rather a human uses a varied collection of inputs and a large number of automated internal parameters that together form the "neural net" architecture. Then "training" occurs via this being a computationally intensive iterative process (often running for days) that explores the interaction between internal variables via stochastic control procedures involving perturbing and monitoring individual variable's effects on overall performance creating a set of "effective" or "optimal" variables. Depending on the specific architecture these parameters number in the range of 10s of thousands to 100s or millions. When "trained" a DB and classifier are automatically generated and transferred to the PPE to run normally . While initially developed with facial recognition tasks in mind this approach has also been successfully repurposed for use with a variety of capital-intensive domains such as detecting cyber threats.

### 5.1. Role of AI in Decision-Making Processes:

Efficient decision-making processes within Security Orchestration, Automation, and Response (SOAR) systems is paramount for effectively managing cybersecurity incidents and mitigating their impact on organizations. SOAR systems have emerged as a solution to address these challenges by integrating and enabling a number of key capabilities in a seamless user experience: orchestration, automation, and decision support for cybersecurity operations. With an overwhelming number of detected security incidents in enterprise environments, a common challenge faced is inefficient and ineffective decision-making, which contributes to delays in responding to vulnerabilities or threats with the highest cyber risks [1].

Artificial Intelligence (AI) offers the promise of more informed, autonomous, and efficient decision-making in SOAR systems through the use of machine learning methods and probabilistic graphical models. Various AI algorithms used in different SOAR systems are explored, including those for predicting the probability of incidents or service tickets leading to successful security events. Other models consider uncertain and probabilistic information to help assess the trustworthiness, risk, and adequacy of the proposed actions for resolving incidents [2]. There is a systematic description of the mathematical models of decision-making processes supported by AI, including the graphical model initiated by a Bayesian Network, a Markov Decision Process, and a hybrid model combining the two paradigms. Lastly, the role of AI in decision support in SOAR systems is discussed, highlighting future directions of research in the field and the challenges in developing AI-based cybersecurity solutions in general.

### 5.2. Enhancing Threat Detection :

SOAR (Security Orchestration, Automation, and Response) systems driven by AI algorithms are elevating threat detection capabilities well beyond the means of more traditional systems by enhancing the proactive measures necessary to identify and respond to potential security threats [2]. The first component of a SOAR platform is the ingestion of data generated from security and IT devices and systems. This data is often inconstant forms and vastly different sets, ranging from raw logs, alerts, and tickets to configuration settings, threat intelligence, and attack patterns, generating a Big Data problem that often overwhelms more traditional SIEM and threat detection platforms [9]. SOAR systems provide mechanisms to harmonize these data inputs through the adoption of a common data model prior to the application of analytics and/or additional categorization and structuring processes. The enriched data is then used to support decisions or drive automated processes.

The growing threat landscape and social climate fostered by the COVID-19 pandemic have introduced new threats, ranging from cyber-physical attacks on infrastructure to the abuse of new technology to amplify the impact of social engineering attacks. Traditionally, front-end threat detection mechanisms such as IDS (Intrusion Detection Systems) and SIEM systems equipped with statistical and rules-based analytics were adopted to avoid information overload and identity relevant or anomalous events within the vast amounts of ingested data. However, modern threats and attack vectors are often sophisticated, polymorphic, and obfuscated, allowing them to effectively bypass such security controls. Advanced analytics and machine learning models can automatically identify patterns, events of interest, or relevant context within a given dataset.

## 6. Challenges and Limitations :

The implementation of automated response solutions raises ethical and privacy concerns, particularly regarding the trade-off between speed and potential collateral damage from automated countermeasures [2]. Adverse implications from a broader perspective, including damage to critical infrastructure or Internet denial, can escalate conflicts or harm innocents. Instead of finding more elaborate retaliations, a focus on shielding and reducing vulnerabilities can be a more prudent approach to avoid cascading unwanted consequences. With the ongoing arms race against tighter cybersecurity, and the malicious employment of AI, a commitment to solutions that extend conflict prevention and de-escalation could improve the current security landscape. Automation levels can drastically change the chain of command and autonomy of response systems, leading to potentially dangerous situations. For example, developing AI-driven offensive capabilities requires robust fail-safes to prevent unwanted attacks inadvertently triggered. The evaluation of those systems also raises serious concerns, especially in regard to responsibility and accountability. At a wider level, the employment of AI in warfare can lead to terrifying scenarios such as the escalation of conflicts, increased difficulty managing the chain of command, or arms escalation with devastating consequences for humankind.

Integration issues represent a complex challenge and involve a twofold solution [10]. On one side, the concern is to include pre-existing security systems in reaction setups, which can be difficult due to issues related to architecture, interoperability, and vendors. On the other side, setting the right parameters behind AI reaction capabilities to effectively balance their advantages and disadvantages can be a challenging and time-consuming effort. The absence of common performance evaluation guidelines to keep track of reaction effectiveness and monitor the performance of those systems in the field exacerbates the problem.

## 7. Conclusion :

As Artificial Intelligence (AI) technology is evolving and permeating multiple social sectors, its usage in cyberspace is increasingly attracting attention. It is perceived as a major game-changer in the combat against cyber threats like hacking, phishing, and various other attacks [2]. At the heart of AI's potentiality in cybersecurity is the promise of automating detection, investigation, and response faster than the attackers can exploit their security holes, and with fewer well-trained security analysts scrutinizing thousands of security alerts every day [1]. However, despite its potentiality, AI awareness is deeply uneven across the existing cybersecurity ecosystem. To unlock AI's potentiality in cybersecurity, the nuances surrounding AI technology must be better understood. An elegant AI modernization strategy that balances the strengths of AI and human expertise is proposed here. First, there is a brief overview of AI's potentiality in enhancing human capabilities through automation, anomaly detection, and advanced decision-support techniques.

Then, there are deeper dives into the domains that remain irreplaceable by AI technology: context understanding, creativity, accountability, intuition, and common sense. Subsequently, there are discussions on human-AI teaming opportunities that can result in a more cybersecurity-resilient ecosystem. Overall, human-AI teaming can significantly improve the way cyber-attacks are mitigated, resulting in a more efficient, resilient, and adaptive cybersecurity ecosystem.

## References:

[1] I. H. Sarker, H. Janicke, N. Mohammad, P. Watters et al., "AI Potentiality and Awareness: A Position Paper from the Perspective of Human-AI Teaming in Cybersecurity," 2023.

[2] S. Bernardez Molina, P. Nespoli, and F. Gómez Mármol, "Tackling Cyberattacks through AI-based Reactive Systems: A Holistic Review and Future Vision," 2023.

[3] G. Srivastava, R. H Jhaveri, S. Bhattacharya, S. Pandya et al., "XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions," 2022

[4] S. Norem, A. E Rice, S. Erwin, R. A Bridges et al., "A Mathematical Framework for Evaluation of SOAR Tools with Limited Survey Data," 2021

[5] C. Islam, M. Ali Babar, and S. Nepal, "A Multi-Vocal Review of Security Orchestration," 2020

[6] R. Kremer, P. N. Wudali, S. Momiyama, T. Araki et al., "IC-SECURE: Intelligent System for Assisting Security Experts in Generating Playbooks for Automated Incident Response," 2023

[7] Z. Tasnim Sworna, C. Islam, and M. Ali Babar, "APIRO: A Framework for Automated Security Tools API Recommendation," 2022

[8] Z. Fan, B. Ghaddar, X. Wang, L. Xing et al., "Artificial Intelligence for Operations Research: Revolutionizing the Operations Research Process," 2024.

[9] M. Schmitt, "Securing the Digital World: Protecting smart infrastructures and digital industries with Artificial Intelligence (AI)-enabled malware and intrusion detection," 2023.

[10]N. Polemi, I. Praça, K. Kioskli, and A. Bécue, "Challenges and efforts in managing AI trustworthiness risks: a state of knowledge," 2024

# Applications of Secure Multi-Party Computation in Financial Services

Brahim Khalil Sedraoui[1], Abdelmadjid Benmachiche[2], Amina Makhlouf[3], and Chaouki Chemam[4]

[1] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*
[2] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*
[3] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*
[4] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*

## Abstract

This article explores the applications of Secure Multi-Party Computation (SMPC) in the financial services sector, highlighting its significance in maintaining data privacy while enabling collaborative computations among multiple parties. As financial transactions often involve sensitive data inputs from various stakeholders, the need for robust security measures is paramount. The paper discusses the challenges associated with data security in financial services, including compliance with industry standards and the complexities of handling large datasets. It also outlines future trends and research directions aimed at enhancing the efficiency of SMPC protocols, making them more viable for extensive applications. The findings suggest that SMPC can facilitate a paradigm shift in financial operations, promoting transparency and privacy in an increasingly digital landscape.

**Keywords**
Secure Multi-Party Computation (SMPC), data security, Privacy-preserving solutions[1]

## 1. Introduction to Secure Multi-Party Computation (SMPC)

Secure multi-party computation (SMPC) enables parties to jointly compute a function while keeping their individual data private [1] It allows a group of untrusted entities to cooperate in a computation without revealing their private inputs to one another. Data can be kept confidential while still being jointly used, resulting in a variety of applications across the financial, health, and marketing sectors. Since it was first proposed in the late 1980s, there have been two noticeable trends in the area of SMPC: a focus on efficient solutions for practical scenarios and on making the solutions usable by non-experts [2]. The range of proposed solutions extends from carefully optimized protocols for a well-defined set of tasks to libraries and frameworks that allow one to transparently instantiate SMPC in generic applications. One of the earliest solutions for actively-secure SMPC. In its basic version, the protocol allows a group of n parties to compute functions on their input's privates 1 ... s n without any party learning anything but the output. This is possible under the assumption of a well-behaved majority, that is more than half of the parties cannot be corrupted. In addition to that, maximal secrecy is achieved in a computational model under the semi-honest adversary model, ensuring that all parties except the outputs of the computation, learn nothing about the inputs.

### 1.1. Definition and Key Concepts

Secure Multi-Party Computation (SMPC) is a technique that allows a group of parties, each with private inputs, to compute a function of their inputs while keeping those inputs private. It allows a group of parties to compute functions over their private data without revealing the data to other parties, which has wide applications ranging from private queries on databases, collaborative filtering, and many other statistical analyses over distributed databases [3].

The basic SMPC model has the following features: each party has its own private input; and the output is given to all parties. Honest majority means that at least half of the parties are honest. The adversary can learn all the information about the inputs of the corrupted parties, such as the reconstruction of corrupted parties' inputs, but no information about honest parties' inputs is given. A computation is statistically secure if no information about the inputs of the honest parties is revealed. A SMPC protocol is called a Yao's solution if each party can either send (commit) to or receive (open) a ciphertext [4]. A polynomial bound on the maximum number of messages is considered as a bounded-round SMPC protocol. In a common reference string (CRS) model, a pre-specified random string, which is only known to parties, is inserted in the computation [5].

### 1.2. Definition and Key Concepts

Secure Multi-Party Computation (SMPC) is a powerful and robust cryptographic protocol that allows multiple parties to jointly compute a function on their inputs while keeping these inputs private [3]. SMPC protocols can provide confidentiality even against passive adversaries who can see all messages exchanged between honest parties but cannot deviate from the protocol. More advanced protocols also consider malicious adversaries who can arbitrarily deviate from the protocol, and they usually rely on expensive zero-knowledge proofs. Still, several recent advances have led to highly practical protocols that enable successful execution of the protocol even in case of some active adversaries. Moreover, there are efficient solutions tolerating up to $t$ out of $n$ actively corrupt parties, with some protocols being computationally efficient under well-established assumptions such as the existence of a homomorphic encryption scheme and secure point functions.

The first financial institutions may have misgivings about outsourcing sensitive data. Therefore, they may first look for solutions for holding sensitive data in common while it cannot reach any party involved in the protocol, the server being only able to perform computations defined by the protocol on these data [2]. If this kind of boxed computation scheme can be trusted, then passive semi-honest secure protocols can be implemented for a few simple classes of functions such as univariate polynomial evaluation, inner products, max, match, or sorting. However, it is known that with one party holding a single security witness, for any arbitrary polynomial computable function, it is impossible to hide the witness from the other party. With richer function classes there is in general no solution, as posited by the result from Yao.

## 2. Importance of Data Security in Financial Services

In the Internet era, the development of online financial services is crucial. Online financial services are the best choice to realize the sharing of financial data and the connection of financial services for the proactive exploration of potential cooperation scenarios. However, because of the continuous increase of sensitive information on the Internet, data security becomes more significant than ever. Due to the complexity and rapid changes of the financial environment, with the tight coupling between online financial services and risk management, multiple security problems arise [2]. Broadly speaking, in the context of financial services, multiple parties, each holding a private input, want to securely evaluate a function of their inputs and share the outcome. The aim is to guarantee that each party learns nothing but the output of the function. The security needs to be guaranteed even when some parties are dishonest.

Data security is a very important issue in financial services. Financial services involve multi-party situation, and every party needs to input its own private data. For example, in loan scenarios, the input of one party may be a credit score, and other parties may input income per month or sales of products. To conduct computations among the multiple parties, the privacy of each party's input should be maintained. However, due to the impossibility of achieving fully secure SMPC in the asynchronous network, the financial services with ideal security cannot be realized directly [3]. In addition, the data to be computed may include millions of transactions, and a high-performance protocol is needed.

### 2.1. Challenges in Data Security

Liabilities stemming from market activity are often collateralized by sensitive data such as security positions and trade-related risk exposures. All in-companies should provide regular off-chain compliance check data to AfB/AuB and Grand Total RAR remotely for an operations-wide independent audit of the CDI and Senior Note AGP Requirements. Also called Parties. Collateral as a liability is a pledge of collateral to a secured party to enhance the attached security interest in: 1. deposit accounts; 2. money; 3. securities; or 4. investment property.

The overall data security landscape in financial services is complex, with 2021 spending on security solutions estimated at billions of dollars. Automated reporting services and monitoring services exist separately, so an SMPC system would have to handle each request to calculate the sub-query result against its database and return it. System design and deployment depend on matching request and return types, allowing single party computation of aggregate numbers or expert witness computations, but other designs may face limitations or data/control loss for manipulated reports [3].

### 2.2. Relevance of Secure Multi-Party Computation

The cornerstone idea of secure multi-party computation (SMPC) is to divide computation into separate pieces, distribute them to distinct participants, and exchange messages so that each party computes the function of private inputs while protecting the privacy of each party's inputs against all the others. SMPC allows distributed computation among mutually distrustful parties on sensitive data, since it enables the parties to jointly compute the desired outputs without revealing anything else. This computational paradigm has prompted great interests in academia and industry, leading to a wide variety of protocols and applications across numerous domains, including database querying [3], intrusion detection, financial analysis, and data mining. However, the efficiency and robustness of such protocols might not meet different required settings of real-world applications, and research on improving the applicability of SMPC is still ongoing.

The financial services sector is one of the primary industries in which sensitive data is more likely to leak. Thus, it is indispensable to preserve the privacy of financial data. Meanwhile, the trend of innovative technologies in financial services, such as data cross-analysis among financial institutions to develop more effective fraud detection strategies, is generating many new opportunities for SMPC. However, the underlying SMPC protocols might not be appropriate for such applications. Hence, it is worthwhile to explore how to adopt existing SMPC protocols to fit into real-world applications or configure new tailored solutions [2]. Recognizing the field of SMPC in financial services helps demonstrate the potential impact of new technology on the financial services industry and serve as a passage for ideas and developments between two disciplines.

## 3. Fundamentals of Secure Multi-Party Computation

### Background

Secure Multi-Party Computation (SMPC), also known as secure function evaluation (SFE), allows a group of parties to jointly compute a public function ($F$) with input values $x_1, x_2, ., x_n$ owned by each party ($P_1, P_2, ., P_{n)}$, while keeping the inputs values private [2]. More formally, the privacy of means that at the end of the computation, every party ($P_i$) learns only its own input value and the computation result, but nothing else.

There are several approaches to SMPC, the most famous being Yao's Garbled Circuits and Goldwasser-Micali-Wicklin (GMW) protocol. In general, SMPC supports secrecy of gives a level of security to computations, though there is a drawback, a vast increase of the complexity of these computations in terms of memory and bandwidth usage. However, there are many applications that are simple enough and at the same time so useful that if done with some security level their impact would be substantial [6].

### Mathematical Preliminaries

Let $G$ be the finite abelian group used as underlying group of the cryptographic scheme. In this context it is assumed $G$ is prime ordered. Let thus $p$ be the prime order of $G$. Let $g$ be a generator of $G$. Given $\alpha\ and\ g,\ G^{\alpha} denotes\ g^{\alpha}$.

In the following, for any $x \in G^{\alpha},\ x_{\{i\}}, i \in \{1, \dots, t\}$ are shares of $x$ such that $x_1 + \dots + x_t = x$, there is the homomorphic operation $(x_1, x_2) \xrightarrow{\oplus} (y_1, y_2)$, where $y_1 = x_1 + x_2\ and\ y_2 = x_2 - x_1$.

Thus, the reconstructing operation

$$(y_1, y_2) \xrightarrow{Dec} x, G^{\alpha} produce\ x\ if\ y_1 + y_2 = x\ .$$

On sharing generation, it is supposed that given ($x \in G^{\alpha}, S_1, \dots, S_t$) are in $\mathbb{Z}_p$ and $y_i = S_i\ mod\ p$, where $i \in \{1, \dots, t\}\ and\ S = S_1 + S_2\ \dots + S_t$, then there will be $S \equiv 0\ (\ mod\ p\ )$.

Given the context, it is assumed that all parties share in a trustedway ( $S_{\{1,P_1\}}, S_{\{2,P_2\}}, \dots, S_{\{t,P_n\}}$).

### 3.1. Mathematical Foundations

Secure multi-party computation (SMPC) can guarantee the joint computation of functions over distributed data without revealing the raw data or the intermediate results to all but authorized parties [3]. In this system, a trusted party like a cryptographic module may be integrated and act as a mediator between parties. Each party inputs a secret into this mediator, which computes a value based on the aggregate of inputs while releasing only the computation result. The trusted party approach can be extended to the model without a trusted party. However, in SMPC, party corruption occurs. Before computing a binary addition, a random number is generated as a secret share. In this scenario, shares of party 1 are ⟨x⟩, ⟨y⟩, while the shares held by party 2 are ⟨x′⟩, ⟨y′⟩.

The following secret-sharing properties are required in SMPC:

(1) Randomness Parameter. The used random number is uniformly distributed in [0,1).

(2) Additive Property. For every party $P_i$, $\langle s\rangle_i+\langle r\rangle_i$ = r.

(3) Secured property. The disclosed information reveals nothing about s to the adversary [5]. Given these properties, a binary addition can be computed. In mathematical terms, the share of the result R equals the sum of each party shares:

$$\langle R\rangle=\langle x\rangle+\langle y\rangle=\langle x\rangle+\langle y\rangle+\langle x'\rangle+\langle y'\rangle=R+\langle r\rangle_1+\langle r\rangle_2.$$

### 3.2. Protocols and Algorithms

Secure Multi-Party Computation is a computation model where data are confidential but distributed among stakeholders, who build a collaborative model without sharing their data. The idea of secure multi-party computation dates back to 1982 when Andrew Yao envisioned a setting in which parties with private inputs want to jointly compute a function of their inputs whilst keeping these inputs private [6].

Protocols for secure multi-party computation can be classified into two main categories, namely, evaluate-and-compile protocols and gate-by-gate protocols. The first category assumes that the parties use a protocol for securely evaluating a function (such as SFE, secret sharing, homomorphic encryption, garbled circuits, etc.) and then compiles the application of other functions (more precisely, other circuits composed of other functions) using this protocol. An example of this approach is to use a protocol for evaluating arithmetic circuits and to compile a digital circuit using adders, multipliers, etc. into its corresponding arithmetic circuit. The second category runs a basic gate protocol for each gate in the circuit [5].

## 4. Use Cases of SMPC in Financial Services

Secure Multi-Party Computation (SMPC) is a cutting-edge cryptographic technique that empowers multiple parties to collaboratively compute a specific function over their respective private inputs while ensuring that these inputs remain confidential throughout the process [7]. This ensures that no party gains access to the sensitive data of others, making SMPC an essential tool for privacy-preserving computations [8].

In the realm of financial services, where sensitive data and strict regulatory requirements intersect, SMPC offers transformative solutions. It allows financial institutions, regulators, and other stakeholders to collaborate securely, enabling advanced analytics, fraud detection, risk assessment, and market insights, all while upholding the highest standards of data privacy and compliance.

SMPC provides a framework for fostering innovation by balancing the need for secure data sharing with the protection of proprietary or personal information. This enables financial organizations to harness the power of shared intelligence without compromising trust or confidentiality. Below are some of the most impactful use cases of SMPC in financial services [9]:

### 4.1. Fraud Detection

Fraud detection is a big part of financial IT services. It is free to accept any transactions, while the firm still need to pay the price whether the transaction is right or not. Thus, it is important to establish security to detect the fraud transactions as soon as possible. SMPC can utilize the advantage of collaboration among different parties. Each party only collects the noise signal due to the consent of all parties unless the big fraud event occurs and the detection power increases dramatically [3].

Financial firm and telecom operator collaboratively detect fraud on telecom network for revenue sharing and on-line context service for mobile user. Leveraging the multi-dimensional information for the transaction and employing the SMPC techniques, the false alarm rate can be controlled while the detection power is preserved. Such computation is feasible with theoretical bound on the computational overhead [7].

### 4.2. Risk Assessment

Risk assessment is one of the important application fields of secure multi-party computation (SMPC). For many financial institutions, the assessment of risk depends on a joint analysis of several portfolios held by different institutions [7]. Such a collaborative assessment of risk is difficult without disclosing each institution's portfolio, and therefore, SMPC offers a possible solution. It has been shown how to compute several popular risk measures, including Value-at-Risk and Expected Shortfall, using SMPC on the underlying risk factors of the portfolios of the institutions. The methods work for fixed-length portfolios and can be extended to handle portfolios with an arbitrary number of instruments. In addition, the SMPC computation can be made more efficient if approximations of risk measures are acceptable [3].

In addition to the joint computation of risk measures, which depends only on the portfolios of the institutions, it would be interesting to conduct a joint analysis of VaR and Expected Shortfall which incorporates additional observable and relevant data (the so-called risk factors) regarding the instruments held on the portfolios. In this context, the risk of a given portfolio and the risk of a group of portfolios are assessed and monitored by the co-innovation of the portfolios with the risk factors. The SMPC can be used to compute these co-variances without revealing the portfolios. It has also been shown how to do this by extending the well-known method of implication on shared secrets to matrix-matrix multiplication [5].

### 4.3. Market Analysis

Within financial services, banks can collaborate in order to create more precise data-driven insights while adequately preserving their sensitive data. For instance, banks can pool their transaction data to obtain accurate estimates concerning the architecture of their payment networks or to determine how consumer or organization behavior differs across banks. This pooled data can also be used to derive safer credit risk estimates, better first-party fraud detection rules, etc. In addition to confidentiality concerns, there are regulatory concerns about excessive data pooling and the potential creation of monopolies or other distorting economic structures. This implies that any pooled data or insights must be secure from leakage or access by non-parties. Cryptographic solutions are required because trusted third parties that do not have access to the pooled data or insights and which are sufficiently powerful to safeguard them, do not exist in practice [7].

Secure Multi-Party Computation (SMPC) enables a set of parties to jointly compute such functions on their private inputs while preserving their confidentiality [3]. This computation takes place in a distributed manner such that at no time do the parties get to know each other's private inputs and where the security of the protocol holds against any coalition of up to at most t-malicious parties [5]. In addition to confidentiality, SMPC also guarantees that the outputs of the function are only revealed to the parties that are entitled to know them. To be practical, the output must preferably be in the form of an aggregated summary that does not permit explicit disclosure of any private input. SMPC has been successfully applied in various domains, such as e-health, auctioning, risk modelling, etc.

## 5. Implementation Challenges and Solutions

The implementation of Secure Multi-Party Computation (SMPC) in financial services is not without its challenges. While the technology holds great promise, there are a number of practical challenges that must be addressed in order for it to be widely adopted in this sensitive and highly-regulated sector [10].

One of the biggest challenges is scalability. Many financial institutions have large numbers of clients and transactions, meaning that any solution needs to be able to handle large volumes of data. Additionally, financial institutions typically operate in a real-time environment, meaning that solutions need to be able to operate quickly and efficiently. Some parties have started exploring off-the-shelf cloud services to implement secure multiparty computation, in particular for (semi-honest) adversaries. This strategy tends to mitigate some security risks, but potentially exposes others. This work addresses some of these challenges and analyzes existing solutions [2].

Another challenge is performance overhead. Privacy-preserving solutions tend to come with performance overhead compared to traditional solutions. There is a risk that some parties opt not to use SMPC solutions simply because they are computationally expensive [6]. There are a number of known techniques to reduce this overhead, such as optimizing the protocol or offloading computation to specialized hardware. Such methods should be explored in parallel with the development of SMPC solutions, to ensure they remain competitive with traditional approaches [10].

Finally, there is the challenge of integration with legacy systems. Financial institutions have typically invested heavily in their operational infrastructure, which tends to consist of a number of disjointed legacy systems. Any new solution must be able to work within these constraints, rather than assuming a complete re-write of the financial services stack. There are known architectural patterns for integrating new systems into legacy architectures, such as Certain architectures. Similar architectural patterns should be considered in the design of SMPC protocol stacks [10].

### 5.1. Scalability Issues

Several use cases addressing the scalability issues of Secure Multi-Party Computation (SMPC) are explored. Deployments of SMPC systems may have varying workloads and computational demands over time, from handling a few hundred transactions a day to thousands of transactions within seconds. Key findings include proposals for a horizontal scalable approach to SMPC, the evaluation of network speed improvements, parallel execution across multiple machines, the use of semi-honest assistant servers, the evaluation of the replicated secret sharing model, and a parallel protocol against malicious adversaries [10].

The potential of SMPC to scale with its underlying network and computational resources is explored, in an effort to allow financial service use cases to protect client's sensitive data in a privacy-preserving manner. The performance of Rounds 2 and 3 of the SMPC protocol is evaluated in isolation. This includes the amount of traffic generated in simulation experiments as well as in real implementations. This evaluation highlights the challenges of achieving high transaction throughput while maintaining optimal network latency. Additionally, proposed solutions to improve the scalability of SMPC are presented, and the results of simulations and implementations are shared. These proposed approaches include batching of transactions at the simple cost of increasing the latency of SMPC, execution of a cascading architecture with multiple networks used in parallel for SMPC, and replication of the same SMPC network across several geographies [3].

### 5.2. Performance Overhead

For use within businesses of third parties, SMPC can have high overhead, since the solution needs to be applied on different services running on different infrastructures; but for the provider, or by setting up a private infrastructure, this is generally a manageable effort. Furthermore, there are several possibilities to reduce the overhead [6], while maintaining security. These possibilities usually come at the price of a more implementation-attracting solution [2]. To reduce the performance overhead of SMPC, especially in a business-to-business context, there are several different strategies that can be pursued.

The first level of abstraction is a business-to-business cooperation planning. Such cooperation should be based on cases of credible baseline trust, levels playing fields, and possible outside auditing capacity. Within such settings, the establishment of minimum operational standards can reduce the necessary adjustments. Another approach to reduce the performance overhead is the efficiency of the protocols applied. Generally spoken, there is a trade-off between the number of messages exchanged and the number of operations conducted [10]. Both components affect the overall time needed to compute the protocol considering the requirements of secure channels, execution time, and number of parties involved.

### 5.3. Integration with Legacy Systems

While the individual and small groups of users typically adopt Secure Multi-Party Computation (SMPC) solutions, global and large institutions usually face the problem of integrating the SMPC with their existing infrastructures. Financial institutions cannot afford to risk any existing services, whether they relate to regulations, competitiveness, Risk Management, or accounting. Moreover, financial institutions are usually subject to regulations covering the handling, storage, and cross-border transfer of customer-provided data. Concentrating data on one party or court might violate such regulations [11].

In such cases, institutions may decide on full SMPC systems, where all computations are conducted in an SMPC manner and, therefore, no plaintext interactions exist [2]. Still, this approach may not be feasible in a grand scheme. The obstacles to multi-party solutions are their complexity and the legal and cultural differences. Additionally, the initial startup costs of hardware and software are another obstacle. In that situation, full point products can still provide significant innovation improvements by deploying them on top of existing single-point solutions. Such systems allow the use of SMPC “a la carte” techniques by integrating them with existing proprietary calculations embedded with the SMPC [11].

As demonstrated on various artificial and real data sets, it is possible to have a general architecture of SMPC-based functions that can be placed on various technologies, regardless of the industrial or technological aspect. These SMPC can be implemented with hardware-based approaches (FPGAs, ASICS) or standard CPUs. Each of such technologies needs a central controller that computes the party-specific information data that will then be uploaded/deployed to the specific party IT system [3]. Such a modular architecture demonstrates its flexibility, with different technologies and vendors meeting a specific party's needs.

If the employment of such architectures is feasible with fixed single-point functions, the SMPC-supplied functions are highly diverse and typically undisclosed. In such a case, the SMPC with its corresponding data and code has to be uploaded to the interested party’s inner circle. The deployment of such complex SMPC cannot remain unattended since this deployment affects the controlling of corporate data and the accuracy of very delicate computations [11]. Therefore, an in-depth understanding of the SMPC and how they will integrate with existing hard- and software components is necessary to overcome the “fear of the unknown” barrier.

## 6. Regulatory and Compliance Considerations

The regulatory and compliance aspects of employing Secure Multi-Party Computation (SMPC) in financial services are significant in that most multinationals must comply with various data protection regulations, such as GDPR and CCPA. In addition, banks and financial services institutions

(FSI) will usually be members of industry guilds and will adhere to specific industry standards. In any scenario where SMPC is used, the FSI must ensure that it is in compliance with industry regulations and standards.

Even with full SMPC protection, there are risks associated with sharing data in the cloud, and organizations must devise and adhere to strict policies to minimize those risks. The solution should be deployed and implemented using guidelines for cloud data protection & encryption, access control (NIST SP 800-162), data segregation, and security incident & event management (SIEM) [3]. As data protection laws apply to any entity that deals with data subjects, it is vital to adhere to the local regulations of any jurisdiction where data subjects reside.

### 6.1. Data Privacy Regulations

Data privacy standards have gained significant awareness and support globally within the last decade. All businesses, especially those dealing with personal identification information, are expected to abide by these ever-evolving regulations and standards. SMPC, being a privacy-preserving paradigm of computation, needs to align its practices with existing privacy requirements and regulations [3]. This includes addressing issues such as the treatment of secure input data, the rights to engage in private computation, how the computed output is treated post-computation, and restrictions on reusing input or output data. The European Union General Data Protection Regulation (GDPR), focused on data protection and consumer privacy, has been one of the main drivers for such initiatives within Europe. In addition to this, each country within the EU has an independent Data Protection Authority. This gives rise to the concern of multi-jurisdictional compliance for MNCs participating in cross-border data founded computations involving companies/clients from multiple countries.

Health-related data security in Brazil, covered under the Brazilian legal framework on health data protection set by the Lei Geral de Proteção de Dados (LGPD), aligns with the 2020 Organization for Economic Cooperation and Development (OECD) recommendation to promote the adoption of trusted, auditable, and privacy-preserving, and compliant technologies like cryptography and differential privacy to facilitate the data handling and transfer for health-related data while adhering to the legal framework [5]. Privacy by design is formally regulated/mandated by the LGPD in the context of security and protection of personal data. It highlights implementing effective privacy-by-design systems as a prerequisite for compliance.

### 6.2. Industry Standards

The implementation of Secure Multi-Party Computation (SMPC) in the financial services sector is subject to existing industry standards, where applicable. These standards exist for major areas such as KYC checks and AML risk monitoring, such as the standards issued by the CSSF in Luxembourg and the European Banking Authority (EBA). Existing industry standards provide guidelines for ensuring compliance and best practices concerning the use of technology and data processing. The implementation of SMPC in financial services should adhere to the relevant industry standards. For example, financial institutions operating in Europe must comply with MiFID II's requirements regarding data handling. Therefore, a suitable SMPC solution must ensure compliance with laws such as MiFID II, Data Protection Law, GDPR, eIDAS, PSD2, CSRD, and alike [2]. Failure to comply with regulations can result in severe fines and legal repercussions. Adherence to industry standards should be considered when implementing SMPC solutions in the financial services sector. Therefore, the SMPC solution architecture must account for existing standards when designing the system. Existing standards can also signify existing and accepted best practices for the use of technology and data processing. Thus, adhering to existing standards can assist in identifying pitfalls in the implementation of new technologies [12].

## 7. Case Studies and Real-World Applications

This section outlines case studies and examples of Secure Multi-Party Computation (SMPC) implemented in the banking sector, insurance sector, and investment management by industry giants such as Enveil, QEDIT, ZKProof, ZKTube, Hive Computing, and others. Examples in product offerings, real-world applications, and academic efforts are included, offering a look at the practical utilization of SMPC across a range of financial contexts.

The Enveil Data Protection platform enables the protection of data in use and discovery at the query level. Users can search, analyze, and collaborate on data without exposing it, using SMPC, homomorphic encryption, and fully homomorphic encryption. Enveil also offers financial data protection services, allowing predictive modeling, risk analysis, and screening for insider trading patterns without sharing raw data [12].

The SMPC protocol by QEDIT provides scalable privacy for on-chain transactions using zero-knowledge proofs (ZKPs). Attesting transactions are possible without exposing sensitive data through zkSNARKs. It can be applied in compliance solutions for transactions between crypto exchanges and financial institutions, protecting customer and transaction data while reducing auditing and compliance costs.

ZKProof is a global initiative promoting ZKP in a standard way. It invites collaboration among academics, organizations, and researchers to drive ZKP research and adoption. Proposed use cases span identity, regulatory compliance, voting systems, and cryptocurrency.

ZKTube aims to provide concrete application examples, tutorials, and a dedicated ZK product vision for Zoom, incorporating products like zkChain, zkAssets, zkID, and zkExchange.

Hive Computing offers a private equity benchmark shard service using AdvanceIQ to mitigate front-running and collusion risks. AMMs compare SHH and RPC in asset pricing accuracy in the financial context, addressing transaction privacy and information asymmetry [2].

## 7.1. Banking Sector

Financial services is one of the most active application domains of SMPC. The financial services sector spends annually about a 100 billion USD protecting itself against cybercrime, fraud and data leakage [3]. Combinational (combinatory) analysis and risk assessment in the banking sector with the SMPC permits different banks to collaborate securely and simultaneously compute credit ratings, risk exposure, and loan defaults, without the bank revealing their own sensitive data to their competitors. The SMPC also permits the financial institution to deal privately with each single budget and execute risk and return analysis over the combination of portfolios without revealing sensitive info of the portfolios and the budget. First real-world applications got developed from Mitre Corporation and the FBI to detect fraud on a nationwide level and from Enron Corporation and GE to assist the NASDAQ Stock Exchange to compute the leading stock in its stock exchanges [2].

## 7.2. Insurance Industry

Insurance business relies on many sensitive data of parties (e.g., customer, insurance companies, clinics, hospitals, etc.) to cooperate but unwilling to share the information, such as estimate insurance premium, detect insurance fraud, etc. Secure Multi-Party Computation (SMPC) enables diverse parties to construct a system on their information without disclosing their information, thus supporting a wide range of insurance applications. A case on insurance premium assessment is constructed as follows. In the insurance industry, the asymmetric knowledge of insurance parties creates various risks. One is the adverse selection problem, i.e., customers hold information that insurers cannot obtain, which causes the imbalance of the insurance pool. Under this situation, a fair and accurate insurance premium may be hard to assess. Generally, expected loss given a customer can reflect the potential risk, where customer information is required. However, according to the insurance principle, the expected loss is closely associated with the customer's "defect" inclination. Therefore, insurance companies are not willing to provide parameters directly (e.g., risk events). The better method is that insurance companies share only the result, i.e., , which does not involve input data. This kind of information-exchange is still highly sensitive since many parties may collaborate to leak extra information. Unfortunately, traditional methods cannot satisfy it. SMPC guarantees that no party can deduce any information on the private input other than the final output.

## 7.3. Investment Management

Most investment strategies depend on research and insight from large pools of data. This data is refined to create a distinct investment thesis with a detailed investment strategy with built-in risk tolerance; in the context of hedge funds, this is referred to as an investment strategy. The investment strategy by hedge funds is controlled through investing and maintaining the proper portfolio of assets; in turn, this is a complex process. Actively maintaining the investments is done using several complex mathematical models, and requires careful consideration of ecological and current news sentiment and other indicators. Using mathematical models entails storing highly sensitive and commercially valuable data such as the resulting equations of analysis and calculations performed on the raw data. There is a framework called the Investment management within the financial services sector, where capital is either invested or loaned to financial markets or enterprises with an expectation of obtaining additional returns; it is generally understood that investments entail risk [3].

Here, Multi-Party Computation (MPC) is applied to investment management, enabling the sensitive information to be kept confidential and event knowledge of the algorithm used is not disclosed to any one party [5]. In the investments case the distribution follows a multitude of hedging and speculation techniques on capital, commodity, and currency markets. There is no ownership of the data but on control over how this data can be used and what properties this data has. Speculation or hedging are economically best efforts, meaning there is an inherent risk of loss. The disadvantage of using MPC constructions for investments is that unless certain needs are met by all parties, the possible loss of capital is equal to the initial investment made by the most trustworthy and dishonest parties combined.

## 8. Future Trends and Research Directions

As SMPC emerges as a powerful tool for privacy maintenance, there is an unequivocal need for continued enhancements in SMPC protocols to make them viable for a wider range of applications. One promising area for future work is improving the efficiency of SMPC protocols to allow the resolution of computation-intensive tasks, such as analyzing large datasets, using SMPC. Currently, the majority of SMPC protocols have complexities that grow linearly with the size of the input. This has made large-scale SMPC computations impractical as the data size increases. Therefore, it is critical for the research community to explore potential solutions for protocols with sub-linear complexities [3]. Furthermore, investigations to enable SMPC to utilize available computing resources more effectively would be beneficial. Such work could involve algorithm designs that allow one SMPC node to provide multiple expensive operations of a function while ensuring the privacy of that function.

There is also scope for future work on allowing the interoperability of SMPC with other security technologies. Cybersecurity technologies are rarely used in isolation, and local areas often implement multiple protection measures simultaneously. Therefore, research on the efficient combination of SMPC with other technologies is both useful and important. A potential design space for such work involves creating a hybrid architecture encompassing different well-established cybersecurity measures while ensuring aggregated security bounds [2]. This would guarantee that the use of SMPC in conjunction with existing technologies (e.g., secure hardware or encryption) would not enable any new attacks on the data and computations when compared to not using the SMPC-enabled technology at all.

### 8.1. Enhancements in SMPC Protocols

An exploration of the future enhancements in Secure Multi-Party Computation (SMPC) protocols is the focus here, envisioning advancements that can further strengthen the security and efficiency of SMPC for diverse financial applications. The envisioned enhancements can be predicted based on an argumentation scheme that involves an enhancement itself followed by one or more beneficial outcomes of that enhancement. The relevant enhancements and outcomes for SMPC protocols to be considered include those focused on efficiency, security, fairness, inclusion, and anonymity.

Future enhancements in SMPC protocols envision engineering protocols that can keep the number of messages low. Communication complexity is one of the heavy elements in protocol evaluation. Thus, protocols with low communication complexity have lower evaluation execution time than those with high communication complexity. This enhancement would entail significant research changes in protocols, which could make SMPC widely applicable for big data applications [6]. Efficient architectures for the execution of SMPC protocols—like using hardware instead of software or applying a design that enables parallel execution and using accelerators—are other better options.

### 8.2. Interoperability with other Security Technologies

There is a significant potential for interoperability with other security technologies. Secure Multi-Party Computation (SMPC) has been extensively studied as a complement to existing Machine Learning (ML) data protection technologies such as Federated Learning (FL) and Differential Privacy (DP) [2]. Future research in this area should focus on studying conditions under which the protocols can be combined. For example, it would be interesting to compute a FL aggregation or update using SMPC and then add DP noise (or vice-versa) and examine the performance tradeoffs. The combination of SMPC and homomorphic encryption may be another interesting avenue [3]. One possible combination of the two would be to use homomorphic encryption on the data and then compute SMPC functions on the encrypted data, taking care to ensure that the encryption can handle the SMPC transformations. If both the above couplets can be combined, it may well be that a three-way combination of all three will be highly synergistic.

Predictive Model Markup Language (PMML) has been widely adopted by the financial services community as a standard for model interchange. It is a future research direction to define and develop an SMPC extension to PMML so that predictive models encoded in PMML can be securely executed in a multi-party environment with guaranteed privacy of the data inputs, predictions and potentially even the model itself [13]. Other open research problems arise from the examples which have been implemented. In the security frameworks for credit scoring and joint customer acquisition, it is assumed that the reliability and integrity of the operating parties (such as the data owner at rest or the computation facility) is guaranteed. A worthy follow-up is to make these systems less dependent on the trustworthiness of the operating parties by implementing them in an environment where at least one of the parties is not fully trusted. Similarly in the credit scoring example, the data computation model is designed such that it can only be used in the context of credit scoring. However, there are many different types of decisions made by different types of industries based purely on the values of some features. These models might look different but have the same data computation paradigm, and thus could also benefit from this technique.

## 9. Conclusion and Summary

The finance sector is moving into a new and more vibrant era that will open up doors to new fields of opportunity and discoverability. The operations within financial institutions will not only become more transparent but also more privacy-preserving. The finance sector must undertake a paradigm shift in the transactions involving individual consumers' preferences and details [12]. Smart contracts have the potential to lead the way and open up a new world of decentralized finance (DeFi). Institutions would enter a new world of transparency and self-enforcement where obligations are digital and do not need any human intervention. However, this comes at the expense of privacy. Most financial products are not standardized and therefore rely heavily on privately disclosed information.

Secure Multi-Party Computation (SMPC) is a family of protocols that become universal when public encryption is added to them. SMPC is fast, efficient, and capable of handling computationally extensive applications [1]. Strategic and high-value SMPC protocols have significant potential applications in the retail finance sector for stock trading, real-time pricing, and risk management, among others, greatly benefiting both the consumers and businesses of the sector. With relatively simple implementations, SMPC protocols could be used to generate more nuanced and granular financial products with fewer systemic risks and even offer solutions to the cry of the current subprime mortgage crisis, including novel carbon trading systems.

## References

[1] M. Rahaman, V. Arya, S. M. Orozco, and P. Pappachan, “Secure Multi-Party Computation (SMPC) Protocols and Privacy,” in *Innovations in Modern Cryptography*, IGI Global, 2024, pp. 190–214. Accessed: Dec. 01, 2024. [Online]. Available: https://www.igi-global.com/chapter/secure-multi-party-computation-smpc-protocols-and-privacy/354040

[2] Z. Ni and R. Wang, “Performance Evaluation of Secure Multi-party Computation on Heterogeneous Nodes,” Apr. 23, 2020, *arXiv*: arXiv:2004.10926. doi: 10.48550/arXiv.2004.10926.

[3] W. Du and M. J. Atallah, “Secure multi-party computation problems and their applications: a review and open problems,” in *Proceedings of the 2001 workshop on New security paradigms*, Cloudcroft New Mexico: ACM, Sep. 2001, pp. 13–22. doi: 10.1145/508171.508174.

[4] I. Zhou, F. Tofigh, M. Piccardi, M. Abolhasan, D. Franklin, and J. Lipman, “Secure Multi-Party Computation for Machine Learning: A Survey,” *IEEE Access*, 2024, Accessed: Dec. 01, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10498135/

[5] C. Guo, A. Hannun, B. Knott, L. van der Maaten, M. Tygert, and R. Zhu, “Secure multiparty computations in floating-point arithmetic,” *Information and Inference: A Journal of the IMA*, vol. 11, no. 1, pp. 103–135, 2022.

[6] F. K. Loy, *Secure Computation Towards Practical Applications*. Columbia University, 2016. Accessed: Dec. 01, 2024. [Online]. Available: https://search.proquest.com/openview/d60b9f066d7899dbda0447b5a9395926/1?pq-origsite=gscholar&cbl=18750

[7] H. A. Javaid, “Improving Fraud Detection and Risk Assessment in Financial Service using Predictive Analytics and Data Mining,” *Integrated Journal of Science and Technology*, vol. 1, no. 8, 2024, Accessed: Dec. 01, 2024. [Online]. Available: http://ijstindex.com/index.php/ijst/article/view/63

[8] V. Chen, V. Pastro, and M. Raykova, “Secure Computation for Machine Learning With SPDZ,” Jan.

02, 2019, *arXiv*: arXiv:1901.00329. doi: 10.48550/arXiv.1901.00329.

[9] K. Ho, “Factors affecting the decision to develop MPC for Collective action in Financial Fraud Industry,” PhD Thesis, Delft University of Technology, 2022. Accessed: Dec. 01, 2024. [Online]. Available: https://repository.tudelft.nl/file/File_d6b61718-418e-4f0f-b57d-4a9e67fea2ba

[10] C. Zhao *et al.*, “Secure multi-party computation: theory, practice and applications,” *Information Sciences*, vol. 476, pp. 357–372, 2019.

[11] G. Oladimeji, “A Critical Analysis of Foundations, Challenges and Directions for Zero Trust Security in Cloud Environments,” Nov. 09, 2024, *arXiv*: arXiv:2411.06139. doi: 10.48550/arXiv.2411.06139.

[12] S. Shukla, G. Sadashivappa, and D. K. Mishra, “Simulation of Collision Resistant Secure Sum Protocol,” Nov. 28, 2014, *arXiv*: arXiv:1411.7756. doi: 10.48550/arXiv.1411.7756.

[13] S. Xu and W. Zhang, “Knowledge as a service and knowledge breaching,” in *2005 IEEE International Conference on Services Computing (SCC’05) Vol-1*, IEEE, 2005, pp. 87–94. Accessed: Dec. 01, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1531242/

# Collaborative Filtering Models Analysis on Amazon Products Dataset

Maroua Benleulmi[1,*,†] , Ibtissem Gasmi[1,†] and Nabiha Azizi[2,†]

[1] *Computer Science and Applied Mathematics Laboratory, Chadli Bendjedid El Tarf University,El Tarf, Algeria*

[2] *Computer Science Department, Labged Laboratory, Badji Mokhtar Annaba University, Annaba, Algeria*

### Abstract

Recommender systems play a crucial role in personalizing user experiences across various domains, from e-commerce to entertainment. This paper explores the application of three collaborative filtering models—Singular Value Decomposition (SVD), Alternating Least Squares (ALS), and k-Nearest Neighbors (k-NN)—to the Amazon Products dataset. By leveraging distinct algorithms, the objective is to uncover their unique strengths and limitations in addressing challenges like data sparsity, cold start and scalability. Through rigorous evaluation, this study sheds light on the comparative effectiveness of these models and offers insights into optimizing recommendation systems for large-scale datasets. The findings pave the way for future innovations, including hybrid approaches and enhanced contextual modeling, to elevate the quality of personalized recommendations.

### Keywords

Collaborative filtering, Singular Value Decomposition (SVD), Alternating Least Squares (ALS), k-Nearest Neighbors (k-NN), recommender systems, Amazon Products dataset.

## 1. Introduction

Recommender systems are a cornerstone of modern information filtering techniques, designed to provide personalized suggestions to users by analyzing their preferences and behaviors [1]. These systems aim to help users navigate vast amounts of data by identifying items of potential interest, such as products, movies, or articles, thereby improving decision-making and user satisfaction. By leveraging historical interactions, demographic information, or content-based features, recommender systems have evolved to become integral in domains like e-commerce, entertainment, and social media.

In e-commerce, recommender systems are vital for enhancing the shopping experience and boosting sales. They help users discover products they may not have explicitly searched for, offering personalized product recommendations based on purchase history, browsing behavior, or preferences of similar users. For platforms like Amazon, recommender systems increase user engagement by presenting curated lists of "frequently bought together" items, "similar products," and personalized deals, driving customer retention and revenue growth. These systems not only improve customer satisfaction but also address challenges like choice overload and cart abandonment.

Collaborative filtering is a widely used approach in recommender systems that focuses on leveraging user-item interaction data to make recommendations. By analyzing patterns in historical user behaviors, collaborative filtering identifies relationships between users and items without requiring explicit content features. For instance, if two users have similar purchasing habits, the system can recommend items purchased by one user to the other. This technique is popular due to its domain-agnostic nature, as it relies only on interaction data, making it suitable for a wide range of applications

[*] Corresponding author.

[†] These authors contributed equally.

m.benleulmi@univ-eltarf.dz (M. Benleulmi); gasmi-ibtissem@univ-eltarf.dz (I. Gasmi); nabiha.azizi@univ-annaba.dz (N. Azizi)

0009-0000-4310-4676 (M. Benleulmi); 0000-0002-8939-1727 (I. Gasmi); 0000-0002-2713-4792 (N. Azizi)

[2-4]. Building on the strengths of collaborative filtering, recent research has introduced enhancements to address challenges such as data sparsity, cold-start issues, and limited personalization. These advancements include hybrid systems that integrate content-based filtering with clustering techniques for improved accuracy, optimization of collaborative filtering algorithms for better resource efficiency and customer retention, and frameworks that incorporate social connections to enhance recommendation precision, relevance, and diversity [5-7].

It is typically divided into memory-based and model-based techniques [8]. Memory-based methods, like user-based and item-based k-Nearest Neighbors (k-NN), compute similarities between users or items directly from the interaction matrix. These methods are straightforward and interpretable but often struggle with scalability and data sparsity. Model-based techniques, such as Singular Value Decomposition (SVD) and Alternating Least Squares (ALS), use machine learning to learn latent factors that represent users and items, enabling robust predictions even for sparse datasets. Both approaches have their advantages, with memory-based methods excelling in simplicity and model-based methods offering better performance on larger datasets.

The choice of a collaborative filtering model depends on multiple factors, including the dataset's characteristics, computational resources, and the specific objectives of the recommender system. For instance, sparsity in the user-item interaction matrix may necessitate using model-based techniques like SVD or ALS, which are more effective in extracting latent features. Scalability and interpretability also play a crucial role, as some models are better suited for large datasets while others offer more transparent predictions. Researchers often rely on evaluation metrics such as accuracy, recall, and F1 score to determine which model performs best under specific conditions.

In this paper, the challenge of choosing the most effective collaborative filtering model is addressed by comparing the performance of three widely used techniques: SVD, k-NN, and ALS, on the Amazon Products dataset. By testing these models on the same dataset, the objective is to provide an unbiased evaluation of their strengths and weaknesses. This study allows the understanding of how each model performs under identical conditions, offering insights into their suitability for real-world e-commerce applications where sparse and large datasets are prevalent.

The remainder of this paper is structured as follows: Section 2 provides an overview of the related work, highlighting key advancements in collaborative filtering techniques. Section 3 explains the methodology and used models. Section 4 presents the experimental setup and evaluation metrics and provides a comparative analysis of the models. Finally, Section 5 concludes the paper with key findings and suggestions for future research directions.

## 2. Related work

Collaborative filtering methods are broadly categorized into memory-based and model-based techniques. Memory-based methods, like k-NN, rely on similarity measures between users or items, making them interpretable but potentially less scalable. Previous studies have demonstrated that k-NN can be effective for small to medium datasets but may struggle with larger datasets due to sparsity issues.

L. V. Nguyen et al. [9] employed the KNN algorithm as a key component of an adaptive recommendation framework designed to enhance collaborative filtering methods. Their adaptive KNN-based model addresses the cold start problem by clustering users based on past interactions and incorporating user cognition parameters into the similarity metrics. This dynamic approach updates user clusters in real-time, enabling more personalized and accurate recommendations, particularly for new users with limited interaction history. The KNN algorithm calculates the similarity between users using various distance metrics, identifies the K most similar users, and computes a weighted average of their ratings to generate tailored recommendations, significantly improving the system’s effectiveness.

Y. Ariyanto et al. [10] utilized the KNN algorithm as a foundational element of their Demographic-Enhanced Cosine-KNN method for movie recommendation systems. This approach calculates user similarities by combining movie ratings with demographic information, such as age, gender, occupation, and zip code, to capture nuanced user preferences. By evaluating cosine similarities, the algorithm identifies similar users and generates personalized recommendations, addressing challenges like data sparsity and cold-start issues. The method's effectiveness was tested against baseline models using the MovieLens 100K and 1M datasets, showcasing substantial improvements in predictive performance and recommendation accuracy.

On the other hand, model-based approaches, such as matrix factorization techniques, decompose the user-item interaction matrix into latent factors, enabling more scalable and robust recommendations. SVD and ALS are two widely studied matrix factorization methods. SVD has been successfully applied in various recommendation tasks since it reduces dimensionality and captures latent user-item relationships effectively. ALS is an iterative algorithm that alternates between optimizing user and item matrices, which has shown success in handling sparse datasets.

F. Nissa et al. [11] employed SVD as a matrix factorization technique to improve a skincare product recommendation system. This approach involves decomposing the user-item rating matrix into three smaller matrices to uncover latent factors influencing user preferences. By addressing data sparsity through filling in missing ratings with average values and using SVD to predict unknown ratings, the system generates personalized product recommendations. This method enhances the accuracy and relevance of suggestions by leveraging historical user ratings effectively.

S. Hong et al. [12] integrated SVD as a key component of the SVD-AE (Singular Value Decomposition Autoencoder) method to enhance the robustness and efficiency of collaborative filtering in recommendation systems. SVD is applied to perform a low-rank approximation of the user-item interaction matrix, enabling the model to capture essential latent factors while reducing the influence of noise in the data. By combining SVD with an autoencoder framework, SVD-AE achieves a closed-form solution that simplifies training and reduces computational complexity. This approach not only improves recommendation accuracy but also demonstrates strong robustness against noisy interactions, making it an effective choice for building reliable recommender systems.

I. G. A. T. A. Sari et al. [13] employed the Alternating Least Squares (ALS) algorithm to enhance the effectiveness of culinary recommendation systems. The study aimed to address the limitations of traditional collaborative filtering methods, which primarily rely on rating data and fail to capture item-specific features like food flavor, ambiance, and price. By leveraging ALS, the proposed system dynamically incorporated diverse item information, enabling more personalized and accurate recommendations. The scalability and efficiency of ALS make it particularly suitable for handling large and sparse datasets, demonstrating its potential to improve the quality of culinary recommendations.

I. R. Budianto et al. [14] utilized the Alternating Least Squares (ALS) algorithm to improve the personalization and effectiveness of video game recommendation systems. The research compared ALS with other matrix factorization techniques, exploring its capability to analyze user preferences and uncover latent patterns in sparse datasets. The study highlighted ALS's balance between accuracy and scalability, making it a viable choice for large-scale recommendation systems. Its application provides an efficient solution for delivering relevant game suggestions, addressing the growing demands of the diverse and expanding video game market.

## 3. Proposed approach

This study focuses on implementing and comparing three collaborative filtering models: SVD, k-NN, and ALS, using the Amazon Products dataset. This dataset provides user-item interactions that form the basis of our user-item matrix, which is essential for collaborative filtering.

### 3.1. Data preprocessing

The Amazon Products dataset was preprocessed to create a user-item interaction matrix. Missing ratings were replaced with the mean rating of each item to reduce the sparsity of the matrix, allowing each collaborative filtering model to have sufficient data for training.

### 3.2. Collaborative Filtering Models

SVD is a matrix factorization technique that decomposes the interaction matrix into latent factors, capturing hidden patterns in user preferences and item attributes.

The predicted rating of user *u* and item *i* is given by:

$$\hat{r}_{u,i} = U_u \cdot \sum \cdot V_i^T \tag{1}$$

Where Σ is the diagonal matrix of singular values, $U_u$ and $V_i$ are the latent features for user *u* and item *i*, respectively.

k-NN algorithm is used to identify the nearest neighbors of a user based on item preferences, allowing recommendations to be made based on similar users. For a user *u* and item *i*, the predicted rating is given by:

$$\hat{r}_{u,i} = \frac{\sum_{v \in N_k(u)} sim(u,v) \cdot r_{v,i}}{\sum_{v \in N_k(u)} |sim(u,v)|} \tag{2}$$

Where $N_k(u)$ is the set of *k* nearest neighbors of user *u.* And *sim(u,v)* is the similarity between users *u* and *v.*

ALS is another matrix factorization technique, optimized iteratively to minimize the error in predicting user preferences. It is especially suited to handle sparse datasets through its alternating optimization process where the user-item interaction matrix *R* is factorized into two smaller matrices *P* (users × latent factors) and *Q* (items × latent factors). It alternates between fixing *P* to solve for *Q*, and fixing *Q* to solve for *P*, minimizing the reconstruction error. Its objective function is given by:

$$\min_{P,Q} \sum_{(u,i) \in R} (r_{u,i} - P_u Q_i^T)^2 + \lambda(\|P\|^2 + \|Q\|^2) \tag{3}$$

Where:

- $r_{u,i}$: Observed rating for user *u* and item *i*.
- $P_u$: Latent factors for user *u*.
- $Q_i$: Latent factors for item *i*.
- $\lambda$: Regularization parameter to prevent overfitting.

### 3.3. Evaluation Metrics

The models are evaluated based on Accuracy and F1 Score for general recommendation quality. Additionally, for the k-NN model, Recall@k and nDCG@k (Normalized Discounted Cumulative Gain) are used to assess the relevance and ranking quality of the top-N recommendations.

Recall@k measures the proportion of relevant items in the top *k* recommendations. It is given by:

$$Recall@k = \frac{|Relevant\ Items\ \cap\ Recommended@k|}{|Relevant\ Items|} \tag{4}$$

*Relevant Items* represents the total of relevant items for a user, while the *Recommended@k* is the number of items recommended in the top *k.*

nDCG@k evaluates ranking quality by considering the order of relevant items in the top *k* recommendations. It is calculated through these steps:

- Calculating the discounted Cumulative Gain (DCG@k):

$$DCG@K = \sum_{i=1}^{k} \frac{Relevance_i}{\log_2(i+1)} \tag{5}$$

*Relevance$_i$* is the relevance of the item at position *i* (rank position).

- Calculating the Ideal Discounted Cumulative Gain (IDCG@k).

$$IDCG@K = \sum_{i=1}^{k} \frac{Relevance\ _i^{ideal}}{\log_2(i+1)} \tag{6}$$

The relevance here is the optimal order of relevance.

- Finally, normalization:

$$nDCG@K = \frac{DCG@k}{IDCG@k} \tag{7}$$

## 4. Experiments and results

### 4.1. Used Dataset

The research utilizes the Amazon Product Dataset introduced by He et al. [15], comprising 142.8 million product reviews collected between May 1996 and July 2014. This dataset includes comprehensive information such as user reviews (ratings, text, helpfulness votes), product metadata (descriptions, categories, price, brand, and image features), and relational graphs ("also viewed" and "also bought" links). In this study, the dataset was used to construct a user-item interaction matrix, which was then split into 80% for training and 20% for testing. This approach ensures that evaluations are conducted on unseen data, maintaining the integrity of the testing process.

### 4.2. Experimental Setup

Each model (SVD, k-NN, and ALS) was trained on the training dataset, and its performance was evaluated on the test dataset. The models were implemented using Python libraries such as scikit-learn and surprise.

The SVD model uses 100 latent factors and a learning rate of 0.005. While the k-NN model tests different values for the number of neighbors (k), specifically 5, 10, and 20. Moreover, the ALS model is configured with a regularization parameter of 0.1, 10 iterations, and 10 latent factors.

### 4.3. Experiment Result

The results are illustrated in Figures 1, 2, and 3, which show the performance of each model over multiple runs or different values of *k*.

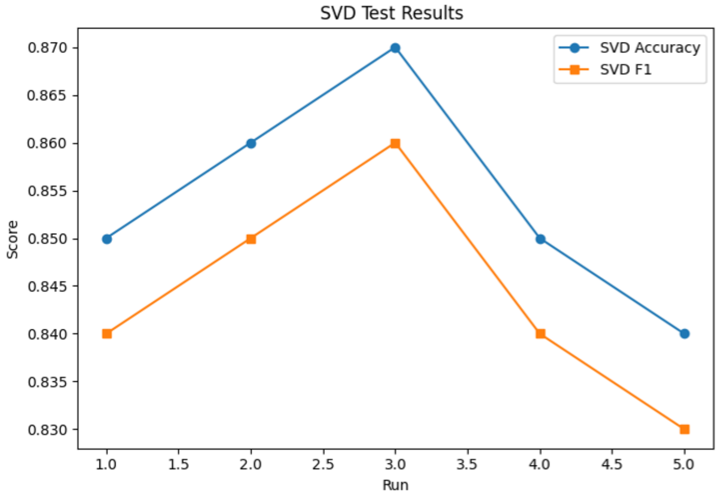


**Figure 1:** SVD test results.

The SVD model demonstrates strong performance, peaking on the third run with an accuracy of 0.87 and an F1 score of 0.855, as illustrated in Figure 1. This peak indicates that SVD successfully identifies and leverages latent factors in the user-item interaction matrix, which enhances the model's ability to predict user preferences accurately. The consistent rise and subsequent plateau in performance over successive runs suggest that the model benefits from parameter tuning, such as adjusting the number of latent factors or learning rate. However, the diminishing improvement after the third run implies that overfitting or saturation in the learning process might occur if further iterations are applied without adjustments. These results emphasize the importance of careful fine-tuning to balance capturing latent patterns and maintaining generalization.

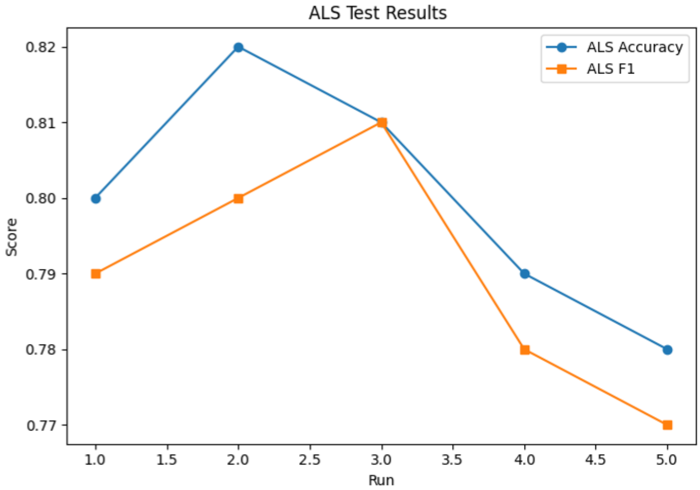


**Figure 2:** ALS test results.

In Figure 2, ALS achieves its best performance during the second run, with a peak accuracy of 0.82 and an F1 score of 0.8. This indicates that ALS effectively handles the sparsity of the dataset and uncovers meaningful latent features. However, the variation in performance across runs suggests that the model's sensitivity to hyperparameters, such as the regularization strength and the number of iterations, significantly impacts its outcomes. The results highlight the need for optimizing these parameters to ensure a balance between avoiding overfitting (regularization too low) and underfitting (regularization too high). While the overall accuracy and F1 scores are slightly lower than those achieved by SVD, ALS's scalability and efficiency make it a competitive option, especially for larger datasets.

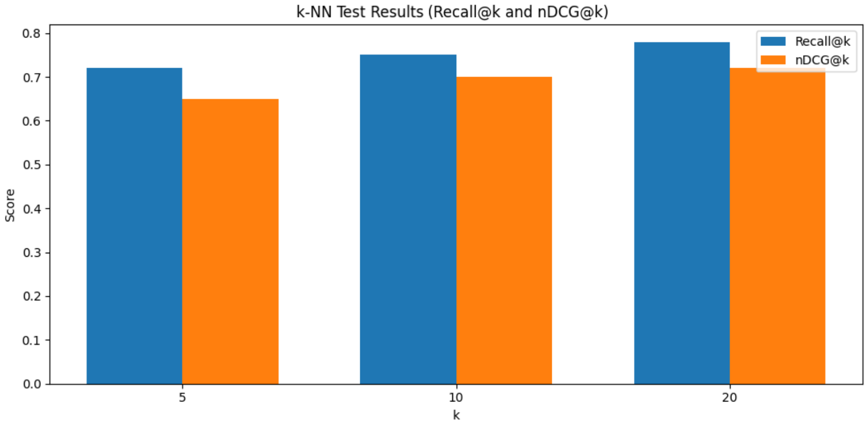


**Figure 3:** K-NN test results.

Figure 3 illustrates the k-NN model's performance based on Recall@k and nDCG@k metrics for k values of 5, 10, and 20. The upward trend in both metrics as k increases indicates that incorporating more neighbors improves the relevance of recommendations and their ranking quality. Specifically, Recall@k values are 0.72, 0.75, and 0.78 for k values of 5, 10, and 20, respectively. Similarly, nDCG@k values are 0.65, 0.70, and 0.72 for k values of 5, 10, and 20, respectively.

Recall@k reflects the system's ability to retrieve relevant items within the top-k recommendations, which increases as k grows, implying enhanced coverage of relevant items. nDCG@k measures the ranking quality of the recommendations, showing that the model positions more relevant items higher as k increases.

However, the diminishing marginal improvements for higher *k* values suggest a trade-off between precision and computational complexity. While larger *k* values improve results, they may also introduce noise from less similar neighbors, reducing the model's ability to focus on truly relevant items. This behavior suggests an optimal range for k, likely between 10 and 20, for balancing recommendation quality and computational efficiency in practical applications.

The k-NN algorithm shows robust performance with an accuracy of 0.78 and an F1 score of 0.77. These metrics highlight the model's ability to provide relevant recommendations to users effectively. The accuracy indicates a solid overall performance, while the F1 score suggests a good balance between precision and recall, meaning the model can correctly identify relevant items without too many false positives or negatives.

### 4.4. Comparison and analysis

SVD outperformed the other models in terms of accuracy and F1 score, achieving peak values of 0.87 and 0.855, respectively. This demonstrates its strong ability to uncover latent factors in the user-item interaction matrix, leading to precise recommendations. However, its performance improvement plateaus after a few iterations, indicating a need for careful parameter tuning to avoid overfitting. While SVD excels in accuracy, its computational cost increases with larger datasets, potentially limiting scalability.

The k-NN algorithm showed robust performance in terms of ranking quality, as indicated by increasing Recall@k and nDCG@k metrics when k values were raised. Additionally, it achieved an accuracy of 0.78 and an F1 score of 0.77. This highlights its capability to provide relevant and contextually appropriate recommendations by leveraging neighborhood-based similarity. However, the gains in performance diminish as k becomes too large, introducing noise from less similar neighbors. Furthermore, k-NN struggles with computational efficiency on larger datasets and sparsity, which can limit its application in large-scale systems.

ALS achieved a balance between scalability and performance, with a peak accuracy of 0.82 and an F1 score of 0.8. Its ability to handle sparse datasets efficiently makes it well-suited for large-scale applications. However, it fell short of SVD's precision, and its performance across runs showed sensitivity to parameter settings, such as regularization and the number of iterations. This suggests that fine-tuning is critical for maximizing ALS's potential.

Among the three models, SVD emerges as the best choice for this study due to its superior accuracy and F1 score, which are critical metrics for recommendation quality. While ALS provides scalability and k-NN excels in ranking quality, the specific requirements of this study—focused on accurate predictions in a relatively controlled dataset size—align closely with SVD's strengths. For scenarios involving much larger datasets or emphasizing ranking over accuracy, ALS or k-NN could be more suitable alternatives.

## 5. Conclusion

This study evaluates the performance of three collaborative filtering models—SVD, ALS, and k-NN—on the Amazon Products dataset to determine their effectiveness in recommendation tasks. Our findings reveal that SVD emerges as the most effective model, achieving the highest accuracy and F1 score, making it particularly well-suited for precise predictions. ALS demonstrates strong scalability and efficiency, offering a compelling alternative for handling large and sparse datasets, while k-NN excels in top-N recommendation quality, showcasing improved Recall@k and nDCG@k scores as *k* increases.

Future work could delve deeper into optimizing each model's hyperparameters through advanced techniques such as grid search or Bayesian optimization to achieve even better performance. Additionally, exploring hybrid recommendation approaches that integrate collaborative filtering with content-based methods or deep learning models could enhance performance, particularly for sparse datasets. Incorporating contextual and temporal features into the models could also offer richer insights and further refine recommendation quality.

## References

[1] Chenguang Pan and Wenxin Li, "Research paper recommendation with topic analysis," 2010 International Conference On Computer Design and Applications, Qinhuangdao, China, 2010, pp. V4-264-V4- 268, doi: 10.1109/ICCDA.2010.5541170.

[2] I. Gasmi, F. Anguel, H. Seridi-Bouchelaghem, N. Azizi, (2021), "Context-Aware Based Evolutionary Collaborative Filtering Algorithm". In: Chikhi S., Amine A., Chaoui A., Saidouni D., Kholladi M. (eds) Modelling and Implementation of Complex Systems. Lecture Notes in Networks and Systems, vol 156, pp. 217– 232, Springer, Cham. https://doi.org/10.1007/978-3-030-58861-8_16

[3] I. Gasmi, H. Seridi-Bouchlaghem, H. Labar, A. Baareh, "Collaborative filtering recommendation based on dynamic changes of user interest", Intelligent Decision Technologies, vol. 9, no. 3, pp. 271-281, 2015. doi:10.3233/IDT-140221

[4] J. Lourenco and A. S. Varde,” Item-Based Collaborative Filtering and Association Rules for a Baseline Recommender in E-Commerce,” 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020, pp. 4636-4645, doi:10.1109/BigData50022.2020.9377807.

[5] T. Pan, "Personalized Recommendation Service in University Libraries using Hybrid Collaborative Filtering Recommendation System," *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)*, Hassan, India, 2024, pp. 1-5, doi: 10.1109/IACIS61494.2024.10721676.

[6] D. Sukhanov, A. Galkin and E. Khabibullina, "Collaborative Filtering Algorithm for Recommender Systems," *2023 5th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, Lipetsk, Russian Federation, 2023, pp. 557-560, doi: 10.1109/SUMMA60232.2023.10349524.

[7] Fareed, A., Hassan, S., Belhaouari, S. B., & Halim, Z. (2023). A collaborative filtering recommendation framework utilizing social networks. Machine Learning With Applications, 14, 100495. https://doi.org/10.1016/j.mlwa.2023.100495.

[8] R. Prabakaran, J. Pradeepkandhasamy and M. Arun, "A Survey on Recommendation Systems using Collaborative Filtering Techniques," *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, 2023, pp. 1445-1450, doi: 10.1109/ICSSIT55814.2023.10060889.

[9] C Nguyen, L. V., Vo, Q., & Nguyen, T. (2023). Adaptive KNN-Based Extended Collaborative Filtering Recommendation Services. *Big Data and Cognitive Computing*, *7*(2), 106. https://doi.org/10.3390/bdcc7020106

[10]D Ariyanto, Yuri & Widiyanigtyas, Triyanna & Ari, Ilham & Zaeni, Ilham. (2024). Enhancing Movie Recommendations: A Demographic-Integrated Cosine-KNN Collaborative Filtering Approach. International Journal of Intelligent Engineering and Systems. 17. 791-803. 10.22266/ijies2024.1231.60.

[11]Nissa, F., Primandari, A. H., & Thalib, A. K. (2023). COLLABORATIVE FILTERING APPROACH: SKINCARE PRODUCT RECOMMENDATION USING SINGULAR VALUE DECOMPOSITION (SVD). *MEDIA STATISTIKA, 15*(2), 139-150. https://doi.org/10.14710/medstat.15.2.139-150

[12]Hong, S., Choi, J., Lee, Y., Kumar, S., & Park, N. (2024). SVD-AE: Simple Autoencoders for Collaborative Filtering. *ArXiv*. https://arxiv.org/abs/2405.04746

[13] E I. G. A. T. A. Sari and Z. K. A. Baizal, "Culinary Recommender System in Yogyakarta Using Alternating Least Squares Collaborative Filtering Method," *2024 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)*, Indonesia, 2024, pp. 327-332, doi: 10.1109/ICoABCD63526.2024.10704410.

[14] F I. R. Budianto, A. Yusrotis Zakiyyah and A. C. Sari, "Web Based Video Games Recommendation System Using Collaborative Filtering Method," *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, Jakarta Selatan, Indonesia, 2023, pp. 549-554, doi: 10.1109/ICIMCIS60089.2023.10348983.

[15] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 507–517. https://doi.org/10.1145/2872427.2883037

# Diagnosis of Induction Machine and Severity Estimation using Novel Gradient Boosting technique [*]

Moutaz Bellah Bentrad[1*], Adel Ghoggal[1] , Tahar Bahi[2] and Seif Eddine Bouziane[3]

[1] *Mohamed Khider University, Electrical Engineering laboratory of Biskra (LGEB) BP 145 RP 07000 Biskra, Algeria.*

[2] *Badji Mokhtar-Annaba University, Electrical Engineering Department, Laboratory of Automation and Signals of Annaba (LASA), P.O Box.12, Annaba, 23000 Algeria.*

[3] *National school of artificial intelligence, Document Electronic Management Laboratory (LabGED),Sidi Abdellah, MVQ8+MF4, Mahelma, Algiers*‹*.*

**Abstract**

This study presents an innovative approach to the diagnosis of induction machine faults through machine learning, using simulated stator current signals acquired from a multi-winding model of an induction machine. Focusing on the detection and classification of air gap eccentricity is a prevalent fault in induction machines. This work distinguishes between static, dynamic, and mixed eccentricities. To ensure robust monitoring capabilities, various severity levels ranging from healthy to early-stage faults up to extreme severity were considered, with simulations conducted under diverse operational conditions. Leveraging a multi-class classification approach, the chosen machine learning model which is a novel approach that combines between gradient boosting and decision trees, demonstrated exceptional accuracy in identifying not only the type of eccentricity but also the severity of each instance, offering a comprehensive diagnostic framework. This approach underscores the potential of machine learning in complex fault diagnosis tasks for induction machines, providing a reliable basis for early detection and targeted maintenance.

**Keywords**

Induction machine, Fault diagnosis, Eccentricity, Machine learning, GBDT.

## 1. Introduction

Fault diagnosis is crucial not only in medical and bioengineering fields but also in industrial applications like motor systems. Among various motor types, induction motors (IMs) remain among the most reliable and widely used, particularly in industries such as electrical, mechanical, and automotive. Their widespread use is due to their numerous advantages, including low maintenance requirements, durable construction, cost-effectiveness, high efficiency, and adaptability under varying load conditions. Additionally, IMs are well-suited for harsh environments, given their robustness. However, continuous and demanding operational conditions can lead to faults that, if left unaddressed, may cause severe failures and additional maintenance costs. Even minor issues can escalate to critical failures that disrupt entire systems. This highlights the growing importance of early fault detection to maintain system reliability, reduce downtime, and extend the operational lifespan of induction motors [1].

[*] Corresponding author.

[†] These authors contributed equally.

moutazbellah.bentrad@univ-biskra.dz (M.B. Bentrad); a.ghoggal@univ-biskra.dz (A. Ghoggal); bahi.tahar@univ-annaba.da (T. Bahi); seifeddine.bouziane@ensia.edu.dz (S. E. Bouziane)

0009-0001-0838-1536 (M. B. Bentrad); 0000-0002-6700-1206 (A. Ghoggal); 0000-0001-8622-2492 (T.Bahi); 0000-0001-5244-1324 (S.E. Bouziane)

Induction motor faults are classified into electrical and mechanical [2]. Electrical faults include overload, open-phase, and short-circuit faults, while mechanical faults include bearing, rotor, and stator faults [3]. Notably, bearing faults are frequent and account for approximately 44% of all faults [4].

Figure 1 illustrates the most common faults in IM:

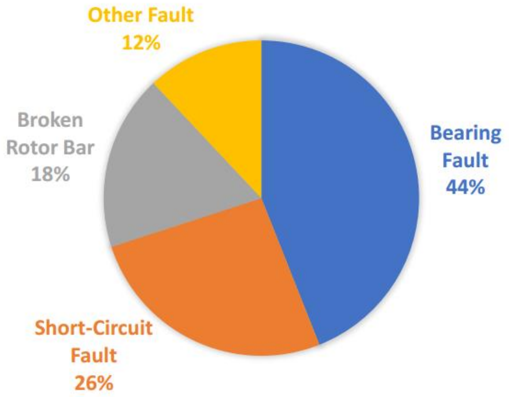


**Figure 1:** Frequent faults of the induction machine.

A variety of artificial intelligence techniques are utilized in predictive maintenance, including Expert Systems (ES), Artificial Neural Networks (ANNs), Fuzzy Logic Systems (FLS), Genetic Algorithms (GAs), and Support Vector Machines (SVM) [5]. Among these, SVM has emerged as a highly popular option due to its strong predictive accuracy and efficiency in reducing both training and testing times, which lightens the computational load during analysis [6]. SVM is particularly effective in diagnosing issues in rotating machinery, especially for faults resulting from excessive vibration. Consequently, it is widely adopted for identifying damage and categorizing different types of faults that can arise in such equipment [7].

In this paper, we aim to detect air gap eccentricity and estimate fault severity in induction machine by classifying multiple samples gathered from simulating the multi-winding model, stator current signal, various types of air gap irregularity (static, dynamic and mixed) with various severities 10%, 20% and 30%. For the classification task a combination between ensemble methods and decision trees called GBDT, the proposed method achieved a great accuracy in terms of classifying each class successfully, which would be helpful for the detection and monitoring of the defect in its early stages.

## 2. Induction machine modeling

Figure 2 represents the rotor of an induction cage machine assimilated to a polyphase winding where each mesh is made up of two adjacent bars and two short-circuit ring portions. The stator circuit is composed of a three-phase winding which can be placed in the stator slots in different ways thus defining the type of winding adopted [8-9]. However, within the framework of the study of the detection and localization of possible faults which occur in the IM, the MWM well proven reliability describes the rotor as a set of loops interconnected between them, each formed by two adjacent bars and which end up with the portions of rings which connect them.

Considering the MWM at the induction cage machine, figures 2 and 3, respectively rotor multi-winding scheme, representation of a rotor mesh from which we can see that each bar and each rotor ring, is modeled by an inductance and a resistance.

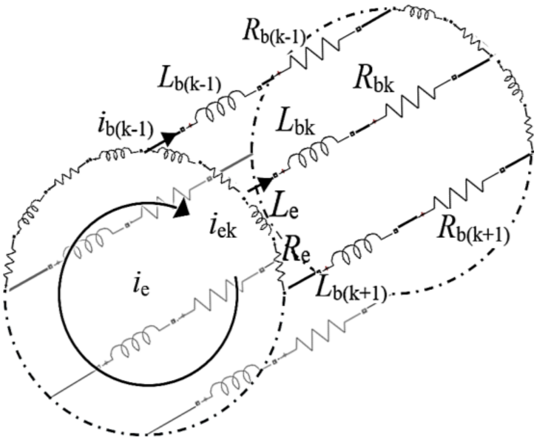


**Figure 2:** Multi-winding model representation

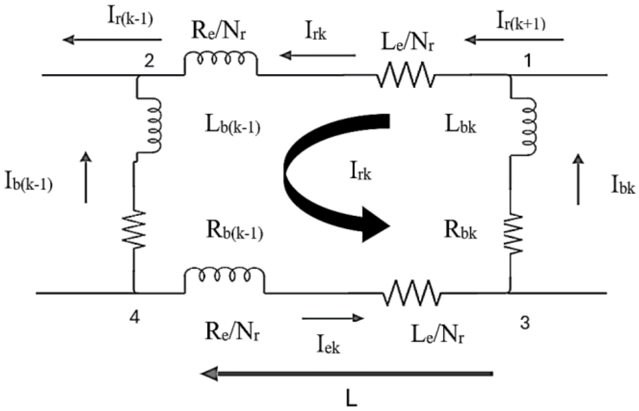


**Figure 3:** One mesh of rotor bar

### 2.1. Voltage mathematical equations

Figure 3 depicts a segment of the equivalent electrical circuit of a rotor mesh, showcasing the rotor bars and short-circuit ring segments through their respective resistances and leakage inductances. Based on this illustration, the equations for the voltages of the three stator phases and the $(N_b + 1)$ rotor meshes can be derived. [10].

$$[V_s] = [R_s][I_s] + \frac{d[\psi_s]}{dt} \tag{1}$$

$$[V_r] = [R_r][I_r] + \frac{d[\psi_r]}{dt} \tag{2}$$

where $[\psi_s]$ and $[\psi_r]$ denote the vectors encapsulating the total fluxes through the stator and rotor windings, respectively. Meanwhile, $[I_s]$and $[I_r]$ represent the respective current vectors associated with these windings.

By joining both of equations (1) and (2) into the same matrix equation, we arrive at:

$$[V] = [R][I] + \frac{d([L],[I])}{dt} \tag{3}$$

Which becomes,

$$[V] = [R][I] + [I] \cdot \frac{d\theta_r}{dt} \cdot \frac{d[L]}{d\theta_r} + [L] \cdot \frac{d[I]}{dt} \tag{4}$$

$$[V] = [R][I] + [I]\Omega_r \cdot \frac{d[L]}{d\theta_r} + [L]\frac{d[I]}{dt} \tag{5}$$

### 2.2. Mechanical equation

Depending on the specific application for which the motor is intended, the mechanical equation of motion can be formulated as follows:

$$J_T \cdot \frac{d\Omega_r}{dt} + f_V \Omega_r = T_e - T_r \tag{6}$$

Where :
$J_T$ : total moment of inertia ;
$f_V$ : viscous friction coefficient ;
$T_e$ : electromagnetic torque ;
$T_r$ : resistant torque.

And after simplifications which finally give the expression of the electromagnetic torque.

$$T_e = \frac{1}{2} \cdot [I_s]^T \cdot \frac{d[L_{sr}]}{d\theta_r} \cdot [I_r] \tag{7}$$

### 2.3. Differential equations

We can group the voltage equations and the mechanical equation into a single matrix representation to arrive at a condensed form [11]:

$$[U] = [B] \cdot [X] + [A] \cdot [\dot{X}] \tag{8}$$

The vector $[\dot{X}]$ is written as follows :

$$[\dot{X}] = [\mathrm{A}]^{-1} \cdot [U] - [\mathrm{A}]^{-1} \cdot [\mathrm{B}] \cdot [X] \tag{9}$$

We thus reveal the state vector $[X]$ and the vector $[U]$ containing the external quantities to the machine such as:

$$[U] = \begin{bmatrix} [V] \\ -T_r \\ 0 \end{bmatrix} \tag{10}$$

$$[X] = \begin{bmatrix} [I] \\ \Omega_r \\ \theta_r \end{bmatrix} \tag{11}$$

## 3. Eccentricty fault overview

Mechanical faults in induction machines typically manifest as eccentricity defects within the air gap[12]. Eccentricity in an electric machine is a gradually evolving phenomenon that can be traced back to the manufacturing stage. During production, various machining and assembly processes can cause misalignment of the rotor with respect to the stator. During operation, two main factors can exacerbate eccentricity. The first is linked to the drivetrain in which the machine operates, where radial forces imposed on the machine's shaft can lead to bearing wear and further misalignment. The second factor, inherent to the machine's operation, is the imbalance caused by eccentricity itself, which disrupts the distribution of radial forces between the stator and rotor [13,14].

Three categories of eccentricity are generally identified [15, 16] :

- **Static eccentricity** (Figure 4(a)) – typically caused by a misalignment of the rotor's axis of rotation with respect to the stator's axis. This is primarily due to improper centering of the end caps.

- **Dynamic eccentricity** (Figure 4(b)) – occurs when the rotor's center of rotation is different from the stator's geometric center, with the rotor's center rotating around this stator center [17]. This type of eccentricity is caused by deformations in the rotor or stator cylinder or by the deterioration of ball bearings.
- **Mixed eccentricity** (Figure 4(c)) – a combination of the two cases described above.

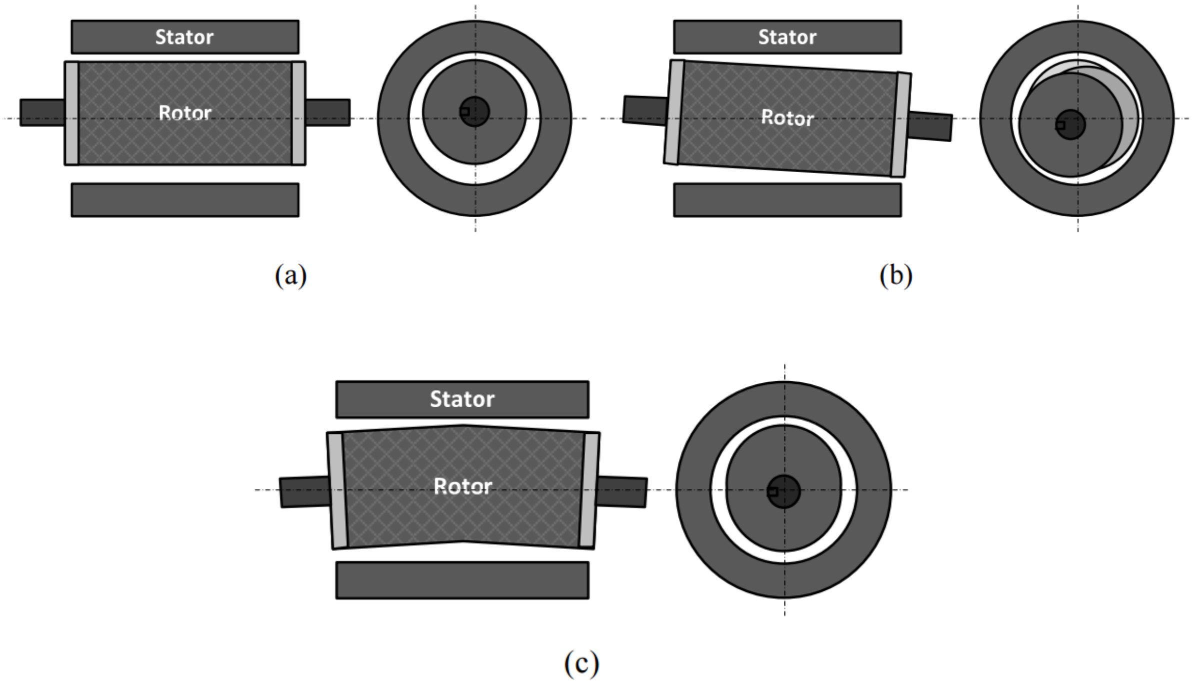


**Figure 4:** Eccentricity types : a-Static, b-Dynamic and c-Mixed [18]

Stator current signals were collected by simulating our model under various operational and severity conditions, figure 5 and 6 shows the impact of eccentricity on the current signal by deforming the sinusoidal form of the current which would result in less efficiency of the machine and more energy absorption.

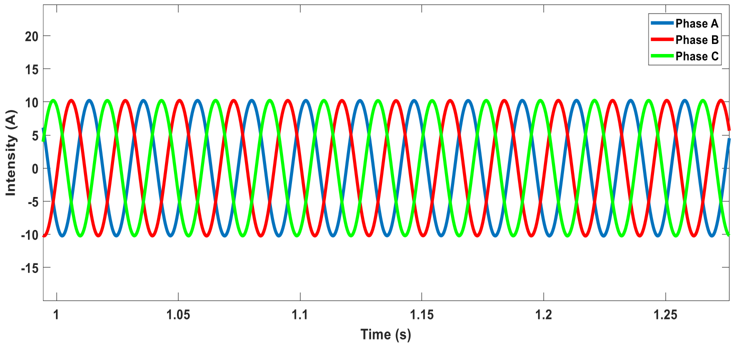


**Figure 5:** Healthy current signal

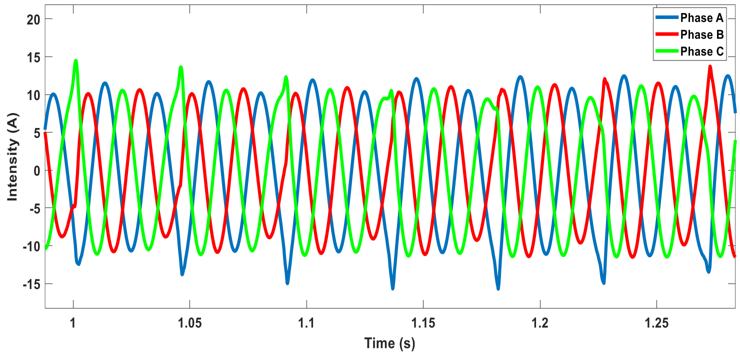
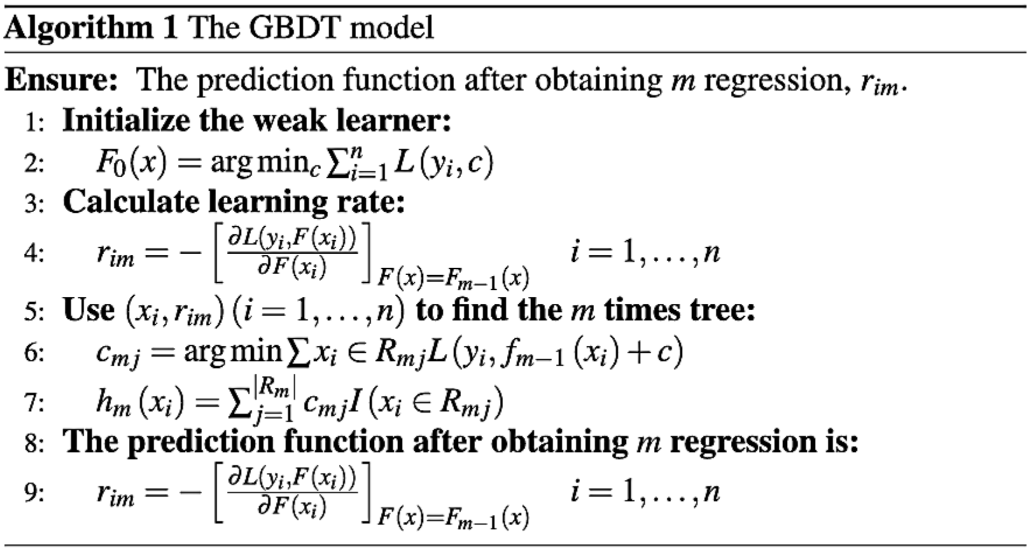


**Figure 6:** Extreme eccentricity current signal

Eccentricty impact on stator current at its early stage can't be visible to the naked eye, which by visual inspection is harder to detect, to metigate this problem more adaptive approach should be considered, which will be discussed in details in the following section.

## 4. GBDT algorithm

GBDT algorithm is based on the combination between Gradient boosting (GB) and Decision Trees (DT), GBDT model can be viewed as a regression model that combines all the weak learners of the model into a strong learner. The following represents the theoretical combination which the model is based on:



```
Algorithm 1 The GBDT model
Ensure: The prediction function after obtaining $m$ regression, $r_{im}$.
Initialize the weak learner:
    $F_0(x) = \arg\min_c \sum_{i=1}^{n} L(y_i, c)$
Calculate learning rate:
    $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad i = 1, \ldots, n$
Use $(x_i, r_{im})$ $(i = 1, \ldots, n)$ to find the $m$ times tree:
    $c_{mj} = \arg\min \sum x_i \in R_{mj} L(y_i, f_{m-1}(x_i) + c)$
    $h_m(x_i) = \sum_{j=1}^{|R_m|} c_{mj} I(x_i \in R_{mj})$
The prediction function after obtaining $m$ regression is:
    $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad i = 1, \ldots, n$
```

**Figure 7:** GBDT model algorithm

GBDT is a gradient boosting method that uses decision trees as its foundational models. It builds an ensemble of weak decision trees by sequentially applying boosting (see Figure 8). In each iteration, a new regression tree is created to minimize the negative gradient of the previous weak learner. The model has an advantage in reducing the risk of overfitting, which can still affect standalone decision trees. Additionally, the proposed model exhibits strong robustness, as it is less sensitive to variations in the size of the training dataset.

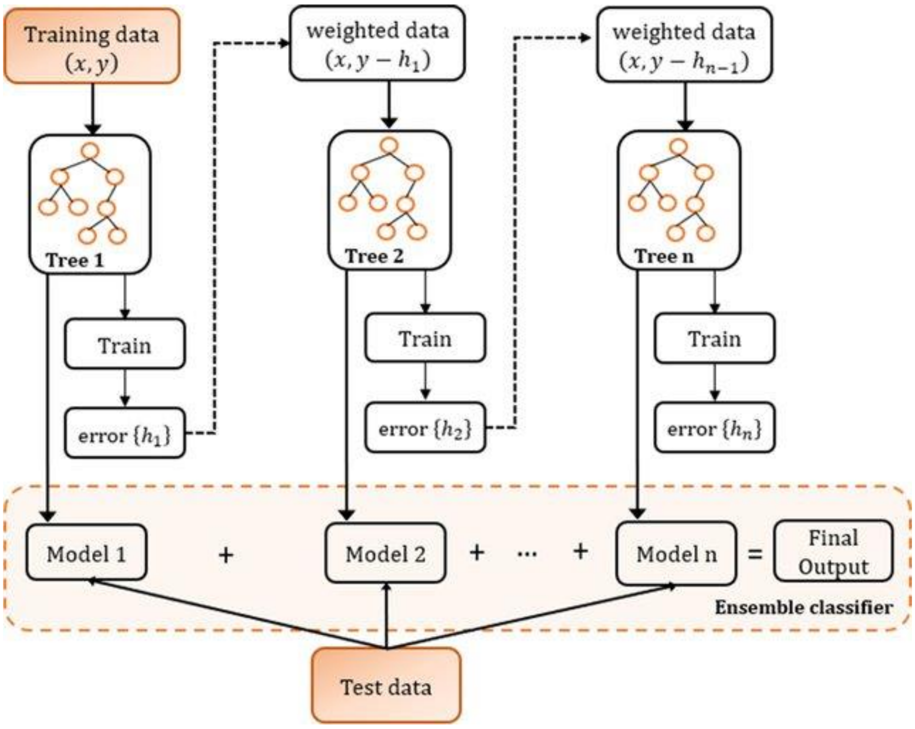


**Figure 8:** GBDT process scheme

The following section is dedicated to showcasing data acquisition and model creation.

## 5. Methodology

Figure 9 shows the process of the proposed model, the model will be able to detect eccentricity defect and classify the defect based on its nature and severity degree ( 10% deemed tolerable, 20% early stage of defect and 30% alarming rate)

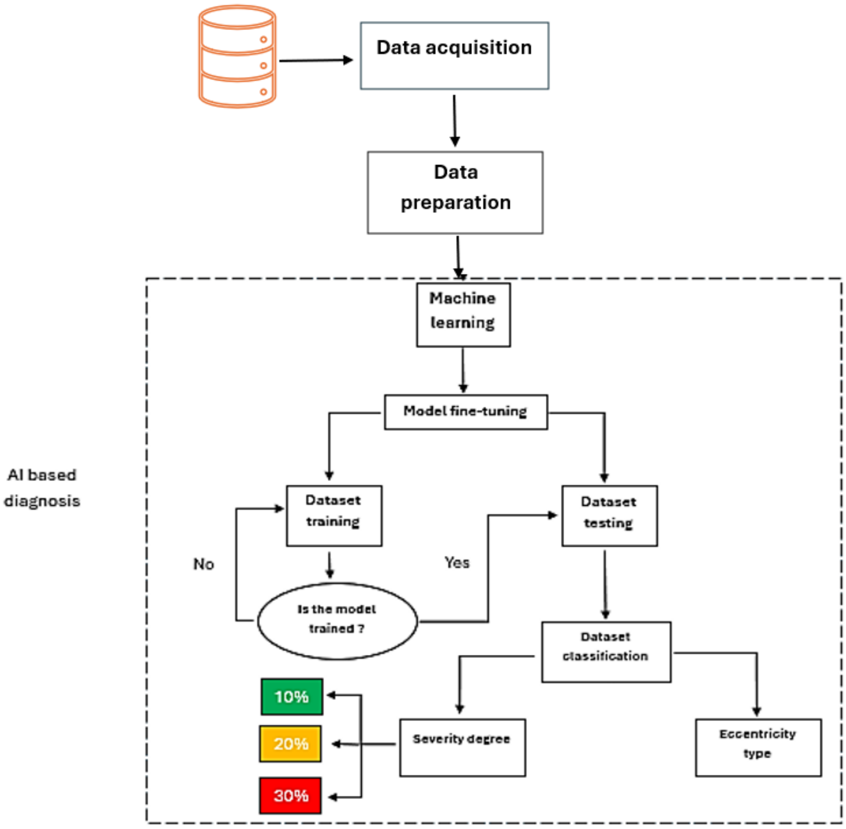


**Figure 9:** Diagnosis methodology

## 5.1. Algorithm application

### 5.1.1. Algorithm architecture :

The chose model was based on the following architecture :

**Table 1:** GBDT architecture

| Component | Configuration |
|---|---|
| Number of classes | 10 |
| Number of trees | 500 |
| Tree depth | 6 |
| Learning rate | 0.05 |
| Max leaves | 31 |
| Feature fraction | 0.8 |
| Bagging frequency | 5 |
| Early stopping | 20 rounds |

These settings create a GBDT model tailored for multi-class fault detection, balancing accuracy with computational efficiency and mitigating overfitting risks. Fine-tuning can be applied based on performance on validation data or cross-validation results, especially if overfitting or underfitting is observed. **is_unbalance** parameter allows GBDT to automatically adjust the data distribution by scaling the positive class weight. When set to true, it helps to balance the classes during training

### 5.1.2. Classification report

The model is evaluated using various metrics such as precision, F1-Score and recall. The proposed model scored an overall accuracy of 97%, evaluating each class individually gave the following results:

**Table 2:** GBDT evaluation

| Class | Precision | recall | F1-Score |
|---|---|---|---|
| Healthy | 1.00 | 0.75 | 0.86 |
| Dynamic20% | 1.00 | 1.00 | 1.00 |
| Dynamic30% | 1.00 | 1.00 | 1.00 |
| Mixed20% | 1.00 | 0.75 | 0.86 |
| Mixed30% | 0.78 | 1.00 | 0.88 |
| Static20% | 1.00 | 1.00 | 1.00 |
| Static30% | 1.00 | 1.00 | 1.00 |
| Macro avg | 0.97 | 0.93 | 0.94 |
| Weighted avg | 0.97 | 0.97 | 0.97 |

The next step consists of applying a 10 folds cross-validation approach to test the model's performance and potential overfitting or underfitting, one can notice in Figure 10 that the model starts with a small training score and then sharply increases in the next iteration with a small provided subset of the actual training dataset thanks to the model's ability to successfully employing weak learners, at the end of training there's a small gap between training accuracy and validation accuracy which suggests that the model has a good generalization capability against new data allowing it to have a good prediction.

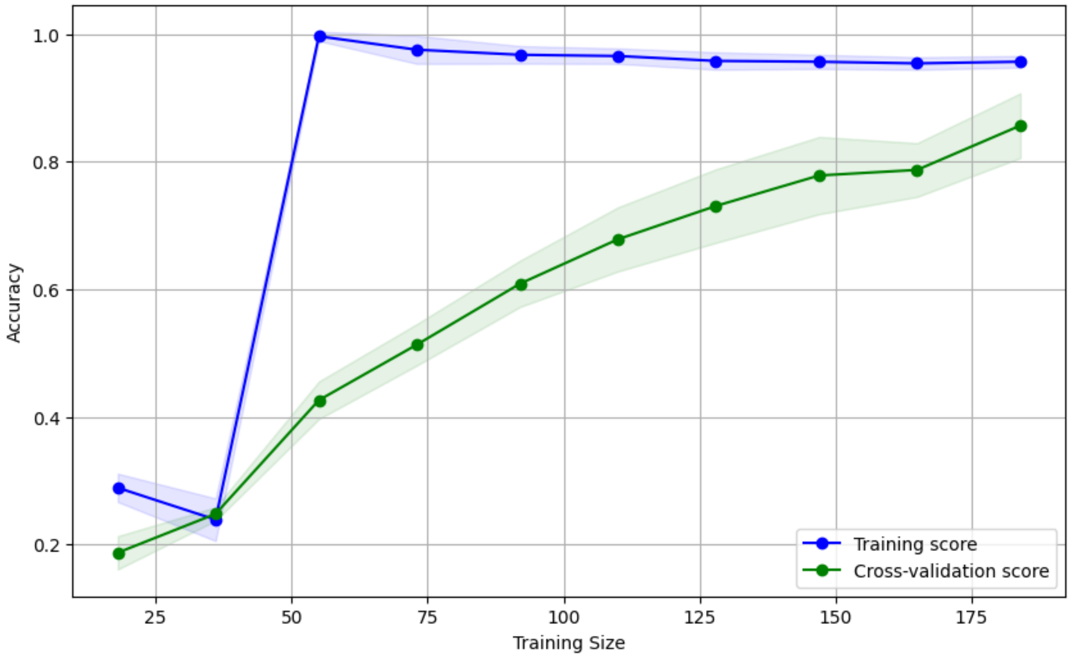


**Figure 10:** learning curve for GBDT model

After ensuring that the model has a good generalization aspect against dataset, another test is conducted to the model's training speed against its original models.

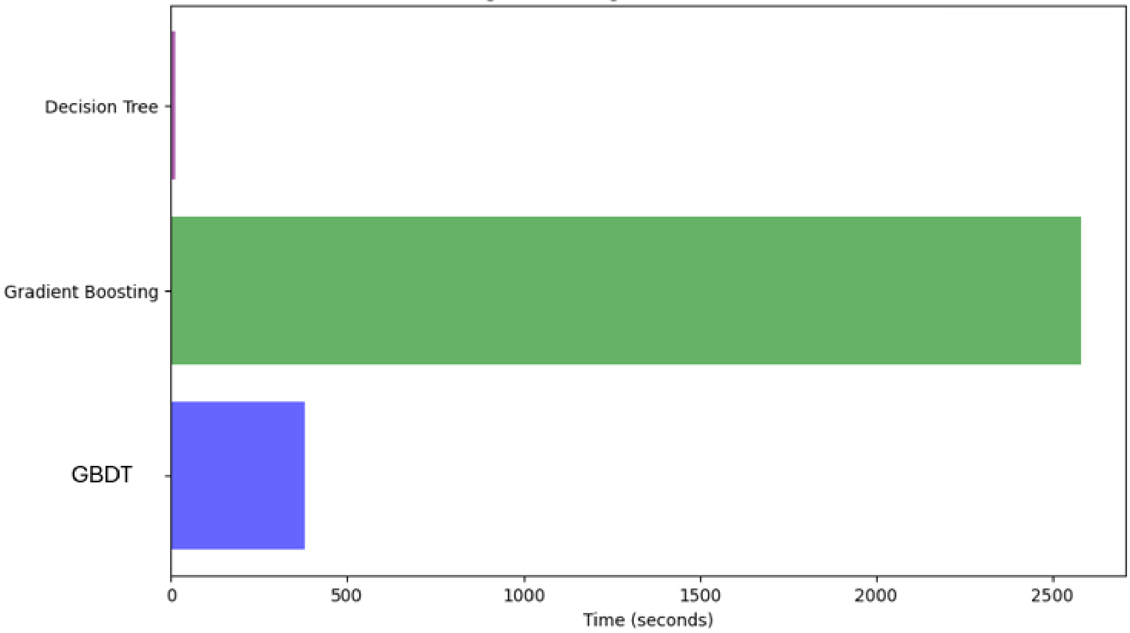


**Figure 11**: Training time for GBDT model against DT and GB

The training time of GBDT comes next after GB. GBDT leverages the boosting technique to combine multiple decision trees, improving performance over a single decision tree. While it is faster than GB, GBDT still involves multiple iterations and calculations, leading to moderate training times. The DT model has the fastest training time among the three. This is because decision trees are straightforward algorithms that partition the data space and build a tree in a single pass. However, this simplicity comes at a cost. Although DTs are quick to train, they often lack the precision of more sophisticated models like GBDT and GB.

In the following part, we will showcase the model's superiority against its original classifiers, the comparison will be conducted under Recall, F1-score and precision to evaluate thoroughly each classifier (see **Table 3**)

**Table 3:** GBDT evaluation against original models.

| Classifier | Recall | Precision | F1-Score | Accuracy |
|---|---|---|---|---|
| DT | 0.84 | 0.85 | 0.84 | 84% |
| GB | 0.97 | 0.95 | 0.96 | 96.5% |
| GBDT | **0.97** | **0.97** | **0.97** | **97%** |

After careful examination GBDT model has a superior performance with a high accuracy compared to its original classifiers which it derives from, indicating that the boosting technique using decision trees proves reliable and fruitful for the task of fault detection.

Figure 12 illustrates the results of the confusion matrix across all

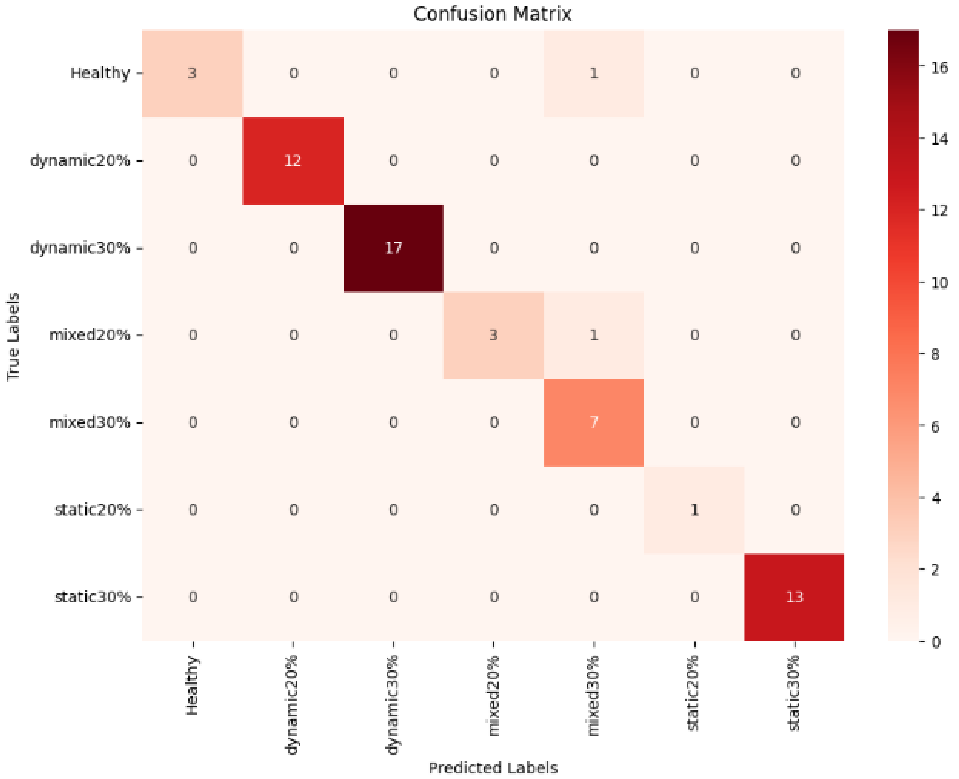


**Figure 12:** Confusion Matrix results

Achieving high accuracy in this context shows that the model is well-tuned and effective at capturing the patterns and features relevant to each class. This performance implies that the model is successfully distinguishing between different types of faults or operating states, even when there are likely overlaps in features or noise in the raw current signal data.

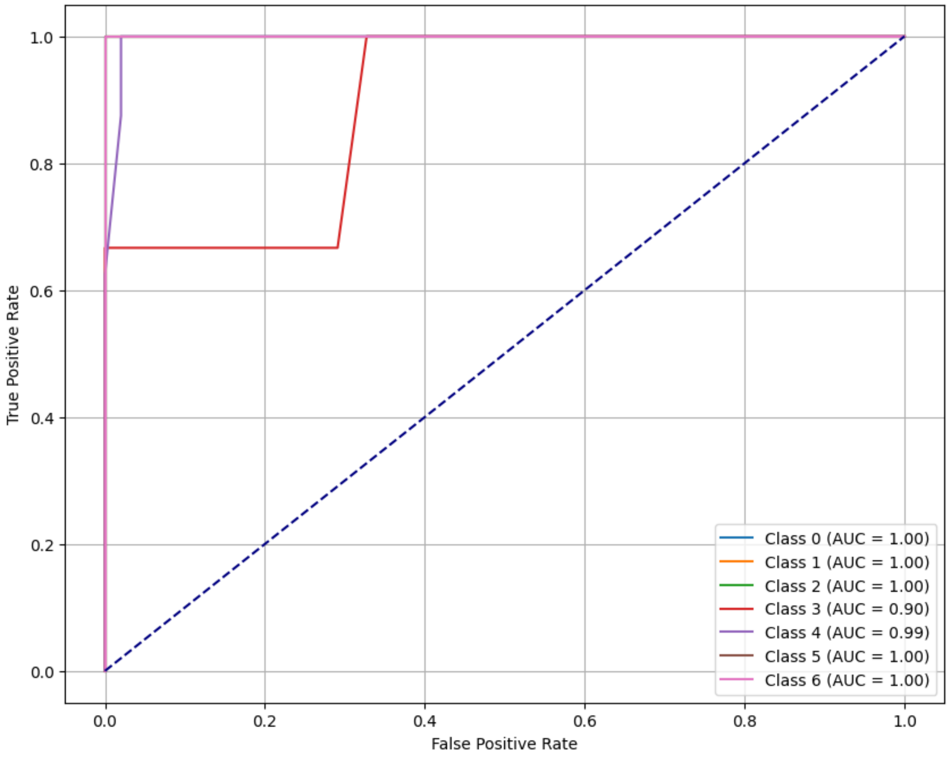


**Figure 13:** ROC curve of the model's performance.

Figure 13 illustrates Receiver Operating Characteristic ,the high Area Under the Curve (AUC) scores across all classes in your ROC curve indicate that your model, using the one-vs-the-rest technique, is performing excellently, providing reliable and precise classification for all the classes involved.

## 6. Conclusion

The model's high level of accuracy demonstrates its robustness and reliability, crucial in real-world applications like fault detection in induction machines, where consistent performance across multiple classes is challenging. Such a result implies that the model effectively captures relevant patterns in the data, even with subtle differences between classes. Given this accuracy, the model is well-suited for practical use, where it could reliably support predictive maintenance by accurately identifying faults, minimizing machine downtime, and helping prevent severe failures.

## Acknowledgements

The authors express their gratitude to the Algerian Ministry of Higher Education and Scientific Research (MESRS) for financially supporting this research.

## References

[1] Yatsugi, Kenichi & Esakimuthu Pandarakone, Shrinathan & Mizuno, Yukio & Nakamura, Hisahide. (2023). Common Diagnosis Approach to Three-Class Induction Motor Faults Using Stator Current Feature and Support Vector Machine. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3254914.

[2] Gangsar, P.; Tiwari, R. comparative investigation of vibration and current monitoring for prediction of mechanical and electrical faults in induction motor based on multiclass-

support vector machine algorithms. Mech. Syst. Signal Process. 2017, 94, 464–481.https://doi.org/10.1016/j.ymssp.2017.03.016.

[3] Misra, S.; Kumar, S.; Sayyad, S.; Bongale, A.; Jadhav, P.; Kotecha, K.; Abraham, A.; Gabralla, L.A. Fault detection in induction motor using time domain and spectral imaging-based transfer learning approach on vibration data. Sensors 2022, 22, 8210.https://doi.org/10.3390/s22218210.

[4] Zarei, J. Induction motors bearing fault detection using pattern recognition techniques. Expert Syst. Appl. 2012, 39, 68–73.https://doi.org/10.1016/j.eswa.2011.06.042.

[5] A. Siddique, G. S. Yadava, and B. Singh, "Applications of artificial intelligence techniques for induction machine stator fault diagnostics: Review," in IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives, SDEMPED 2003 - Proceedings, (2003), pp. 29–34.

[6] P. Gangsar and R. Tiwari, "Comparative investigation of vibration and current monitoring for prediction of mechanical and electrical faults in induction motor based on multiclass-support vector machine algorithms," Mech. Syst. Signal Process., vol. 94, pp. 464–481, (2017)

[7] W. Wanto, R. Lulus. G. H., and D. Djoko Susilo, "Diagnosis Ketidaklurusan (Misalignment) Poros Menggunakan Metode Multiclass Support Vector Machine (Svm)," Mek. Maj. Ilm. Mek., vol. 18, no.2, pp. 39–43, (2019).

[8] L. Heming, S. Liling, X. Boqiang. "Research on transient behaviors and detection methods of stator winding inter-turn short circuit fault in induction motors based on multi-loop mathematical model", In: Proceeding of IEEE ICEMS, 27-29 September 2005; 3: pp: 1951 – 1955.

[9] X. Luo, Y. Liao, H.A. Toliyat, A. El-Antably, and T.A. Lipo, "Multiple coupled circuit modelling of induction machines," IEEE Trans. Indust. Appl., vol. 31, March/April 1995, pp: 311-318.

[10] H. Razik, G. Didier, "Notes de cours sur le diagnostic de la machine asynchrone," Notes de cours, I.U.F.M. de Lorraine, Maxeville, 7 Janvier 2003.

[11] N.Yassa, M.Rachek, H.Houssin. Motor Current Signature Analysis for The Air Gap Eccentricity Detection In The Squirrel Cage Induction Machines. Energy Procedia 162 (2019) 251–262

[12] S. Bazine, "Conception et implémentation d'un Méta-modèle de machines asynchrones en défaut," Thèse de doctorat, Laboratoire d'Automatique et d'Informatique Industrielle (LAII) - EA 1219, Université de Poitiers, 2009.

[13] D. G. Dorrell and A. C. Smith, "Calculation of UMP in induction motors with series or parallel winding connections," IEEE Transactions on Energy Conversion, vol. 9, pp. 304-310, 1994.

[14] D. G. Dorrell, W. T. Thomson, and S. Roach, "Analysis of airgap flux, current, and vibration signals as a function of the combination of static and dynamic airgap eccentricity in 3-phase induction motors," IEEE Transactions on Industry Applications, vol. 33, pp. 24-34, January/February 1997.

[15] S. Nandi, T. C. Ilamparithi, L. Sang Bin, and H. Doosoo, "Detection of Eccentricity Faults in Induction Machines Based on Nameplate Parameters," IEEE Transactions on Industrial Electronics, vol. 58, pp. 1673-1683, May 2011.

[16] R. N. Andriamalala, H. Razik, L. Baghli, and F. M. Sargos, "Eccentricity Fault Diagnosis of a Dual-Stator Winding Induction Machine Drive Considering the Slotting Effects," IEEE Transactions on Industrial Electronics, vol. 55, pp. 4238-4251, December 2008.

[17] H. Razik, "Le contenu spectral du courant absorbe par la machine asynchrone en cas de défaillance, un état de l'art," La revue 3EI, vol. 29, pp. 48-52, Juin 2002.

[18] ANDRIAN CEBAN, thèse 2012 :'METHODE GLOBALE DE DIAGNOSTIC DES MACHINES ELECTRIQUES'

# Early Parkinson's Detection: A Speech-Based Deep Learning Model with LSTM for Accurate Diagnosis

Djaidja Imane[1], Maamri Ramdane[2], Boulmerka Aissa[3] and Harous Saad[4]

[1] *Abdelhafid Boussouf University Center of Mila, Lire Laboratory Abdelhamid Mehri Constantine 2 University, Algeria*

[2] *Abdelhamid Mehri Constantine 2 University, Lire Laboratory Abdelhamid Mehri Constantine 2 University, Algeria*

[3] *The National Higher School of Articial Intelligence (ENSIA) , Algeria*

[4] *College of Computing and Informatics, Department of Computer Science, University of Sharjah, Sharjah, United Arab Emirates*

### Abstract

Parkinson's disease (PD) is a gradual, neurodegenerative condition that often reveals its symptoms only in the later stages. Early detection and intervention can significantly improve the quality of life for those with PD, yet this is hindered by the overlap of PD symptoms with those of other disorders. To address this challenge, our study proposes an innovative PD diagnosis method that harnesses speech signals' power based on deep learning methods. Deep learning, known for its exceptional performance in diverse fields, is utilized here for early-stage PD detection. Specifically, we introduce a novel approach for classifying PD speech datasets, employing a customized Long Short-Term Memory (LSTM) network designed for this purpose. Our model's evaluation on a Parkinson's dataset showcases remarkable results, achieving an impressive F1-score of 0.98 and an AUC of 97%. These findings unequivocally demonstrate the superiority of our focused deep learning approach over current state-of-the-art models in detecting Parkinson's disease using speech signals.

### Keywords

Parkinson's disease, Deep learning, LSTM, Gradient Boosting, Random Forest, AdaBoost

## 1. Introduction

Parkinson's disease (PD) is a progressive neurological disorder that affects millions of people worldwide, leading to significant disability and impairment in daily life. Early detection of PD is crucial for timely intervention and management of symptoms, as current treatments are more effective in the early stages of the disease. Traditional methods of PD diagnosis rely on clinical evaluation, which can be subjective and may not always lead to early detection. Recent advancements in machine learning (ML) and deep learning (DL) have provided new opportunities for the early detection of PD. These techniques allow for the analysis of large datasets, such as speech recordings, to identify patterns and markers that may indicate the presence of the disease. ML and DL algorithms can extract complex features from these datasets, providing insights that may not be apparent to human observers. One promising approach in ML is the use of Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) well-suited for sequence modeling. LSTM networks have been successfully applied in various natural language processing tasks and have shown promise in analyzing sequential data, such as speech signals. By leveraging the power of LSTM networks and other ML techniques, researchers aim to develop more accurate and reliable methods for the early detection of PD. This study investigates the effectiveness of LSTM networks and other ML approaches in detecting PD from speech recordings. We compare the performance of these techniques with traditional clinical evaluation methods to determine their accuracy and efficiency in diagnosing PD. Our goal is to contribute to developing more effective diagnostic tools for PD, ultimately improving patient outcomes and quality of life.

*The 2nd International Conference on Autonomous Systems and their Applications (ICASA'24)*

imen.djaidjaa@centre-univ-mila.dz (D. Imane); ramdane.maamri@univ-constantine2.dz (M. Ramdane); aissa.boulmerka@ensia.edu.dz (B. Aissa); harous@sharjah.ac.ae (H. Saad)

## 2. Related work

In the field of predicting Parkinson's Disease (PD), researchers globally have explored the use of Machine Learning (ML) and Deep Learning (DL) methods. While these approaches hold promise, using multiple models often leads to varying outcomes.
This section provides an overview of some notable approaches in the literature. Abdurrahman et al. used the XGBoost algorithm to prognosticate Parkinson's disease, with an accuracy score of 85.60% [1]. Pérez et al. developed a system to distinguish Parkinson's disease from speech recordings. They used 27 features from sustained vowel recordings and the Hoehn and Yahr scale, achieving an accuracy score 85.25% using SVMs and cross-validation, a significant advancement in medical diagnostics [2]. Moreover, Sakar et al. conducted a comprehensive assessment of signal processing systems used in PD diagnosis from voice. The SVM approach, using the top 50 features highlighted by the mRMR method on the combined subsets, achieved the highest accuracy of 86% [3]. Bocklet et al. identified PD using vocal, acoustic, and prosodic features, collecting data from 46 Czech native speakers (23 with PD) via the Rusz et al. [] study. They examined various speech tasks and used a correlation-based feature selection (CFS) method. Using a support vector machine (SVM) classification algorithm, they achieved an impressive 0,91 accuracy with prosodic modeling, demonstrating the superior performance of prosodic features .[4] Mathur et al. used multiple ensemble approaches to predict Parkinson's disease, including KNN with Adaboost, KNN with Bagging, and KNN with MLP. They found that both KNN with Adaboost and KNN with MLP had the highest accuracy at 91.28% [5]. We proposed a Long-Short-Term Memory (LSTM) deep learning approach to detect Parkinson's disease (PD). The proposed approach surpasses state-of-the-art methods in both F1-score and AUC evaluation metrics. These evaluation metrics are obtained by applying the proposed method to the Parkinson's Dataset.

## 3. Experimental results

### 3.1. Parkinson's Dataset description:

This dataset comprises recordings from 31 patients, both male and female, totaling 195 voice samples. Among these individuals, 23 were diagnosed with Parkinson's disease (PD), while 8 were classified as healthy controls. Each patient provided approximately six recordings, each lasting between 1 to 36 seconds. The primary goal of this dataset is to differentiate between healthy individuals and those with PD. Voice recordings were made using an Industrial Acoustics Company (IAC) AKG-C420 Head-mounted Microphone in a soundproof studio, positioned about eight centimeters from the patient's mouth. The dataset was collected to investigate the diagnostic relevance of Parkinson's disease on speech and vocal irregularities. It can also be utilized to examine the impact of PD on vocal characteristics and the diagnostic significance of vocal symptoms. By including a substantial number of patients at various disease stages, this comprehensive dataset was compiled for analysis. The first column of the dataset contains the names of the patients, providing a solid foundation for further research on PD and its effects on speech [6].

### 3.2. Exploratory Data Analysis:

The key vocal acoustic parameters crucial for medical diagnosis encompass noise measurement, vocal extension, acoustic spectroscopy, and perturbation index. These parameters fall into distinct categories, including pitch, amplitude, voice noise, and nonlinear measures. The dataset comprises 8 individuals in good health and 23 individuals with Parkinson's disease, ensuring a balanced representation. The visual representation of the relationships between these features in Fig.1 enhances the clarity of the dataset description.

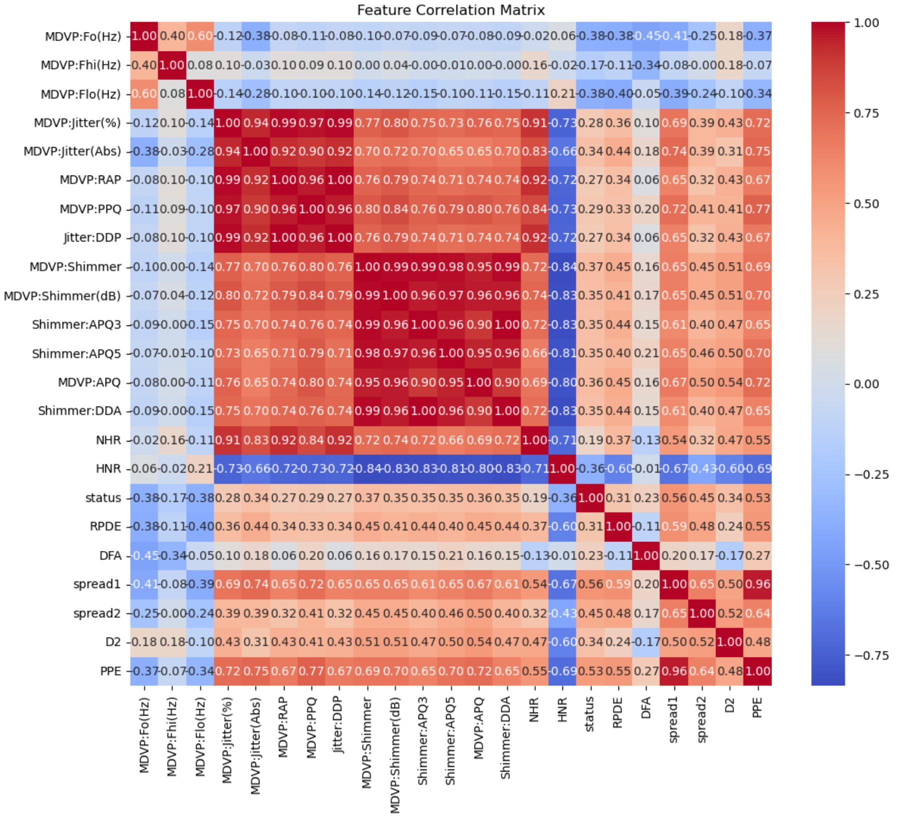


**Figure 1:** Feature Correlation Analysis

# 4. Methodology

Machine learning and Deep learning have demonstrated remarkable performance across various applications, particularly in the early-stage detection of Parkinson's disease. Three machine-learning classifier models and one deep-learning model were used in this paper, and their performances were evaluated.

## 4.1. Machine Learning models

### 4.1.1. Random Forest:

Random Forest is an ensemble learning method that combines the predictions of numerous decision trees to create a robust predictive model. It adds unpredictability to the training process by selecting random subsets of data and characteristics for each tree, resulting in a more resilient and accurate model. The final prediction is produced by combining the forecasts of individual trees, making Random Forest less prone to overfitting and more suited to a variety of machine-learning problems.

### 4.1.2. Gradient boosting:

Gradient boosting is a technique used to build sophisticated models by combining many weaker models. Its goal is to minimize the decrease in performance with each subsequent model. It works by iteratively fitting new models to the errors of the previous models using gradient descent. This approach helps improve the precision of each new model. However, it's important to stop boosting at some point to prevent the model from becoming overfit to the training data.

#### 4.1.3. AdaBoost:

AdaBoost is an ensemble learning technique that sequentially combines the predictions of weak classifiers to create a robust and accurate model. Weak classifiers are typically simple models with performance slightly better than random chance. During training, AdaBoost assigns higher weights to misclassified instances, allowing subsequent weak classifiers to focus more on the challenging examples. The final prediction is then determined by a weighted vote or a weighted sum of the individual weak classifiers. AdaBoost is effective in improving model performance, particularly in situations where a single weak classifier may struggle.

### 4.2. Deep Learning model

#### 4.2.1. LSTM (Long Short-Term Memory):

LSTM is a type of recurrent neural network (RNN) architecture, which was developed to deal with the many long-term dependencies learning issues from the sequential input. It employs memory cells and gating systems to choose only the information that should be stored or erased at specific time spans. LSTMs perform very well at capturing the context and the relationships between the time series or sequence data; hence, they are very suitable for activities like natural language processing, speech recognition, time series prediction, and many others.

## 5. Experimental results:

### 5.1. Performance metrics:

Several metrics are available for evaluating the performance of classifiers in classification tasks. These metrics are based on common terms, including:
**True Positive (TP)**: occurs when both the prediction and the ground truth are positive.
**True Negative (TN)**: occurs when both the prediction and the ground truth are negative.
**False Positive (FP)**: occurs when the prediction is positive, but the ground truth is negative.
**False Negative (FN)**: occurs when the prediction is negative, but the ground truth is positive.
**Precision**: It is the ratio of correctly predicted positive observations to the total predicted positive observations by the classifier, expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

**Recall**: It is the ratio of correctly predicted positive observations to the all observations in the actual class, expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

**F1-score**: It is calculated as the harmonic mean of Recall and Precision, given by:

$$F1 - score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{3}$$

**Accuracy**: It is the ratio of correctly predicted observations to the total observations computed by the classifier, expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

### 5.2. Performance Results and discussion

Table 1 compares various Machine and Deep learning classifiers used in this research. The proposed LSTM model achieves the best accuracy score of 97The comparative analysis of machine learning models for Parkinson's disease (PD) detection reveals interesting findings. Among the models evaluated, the

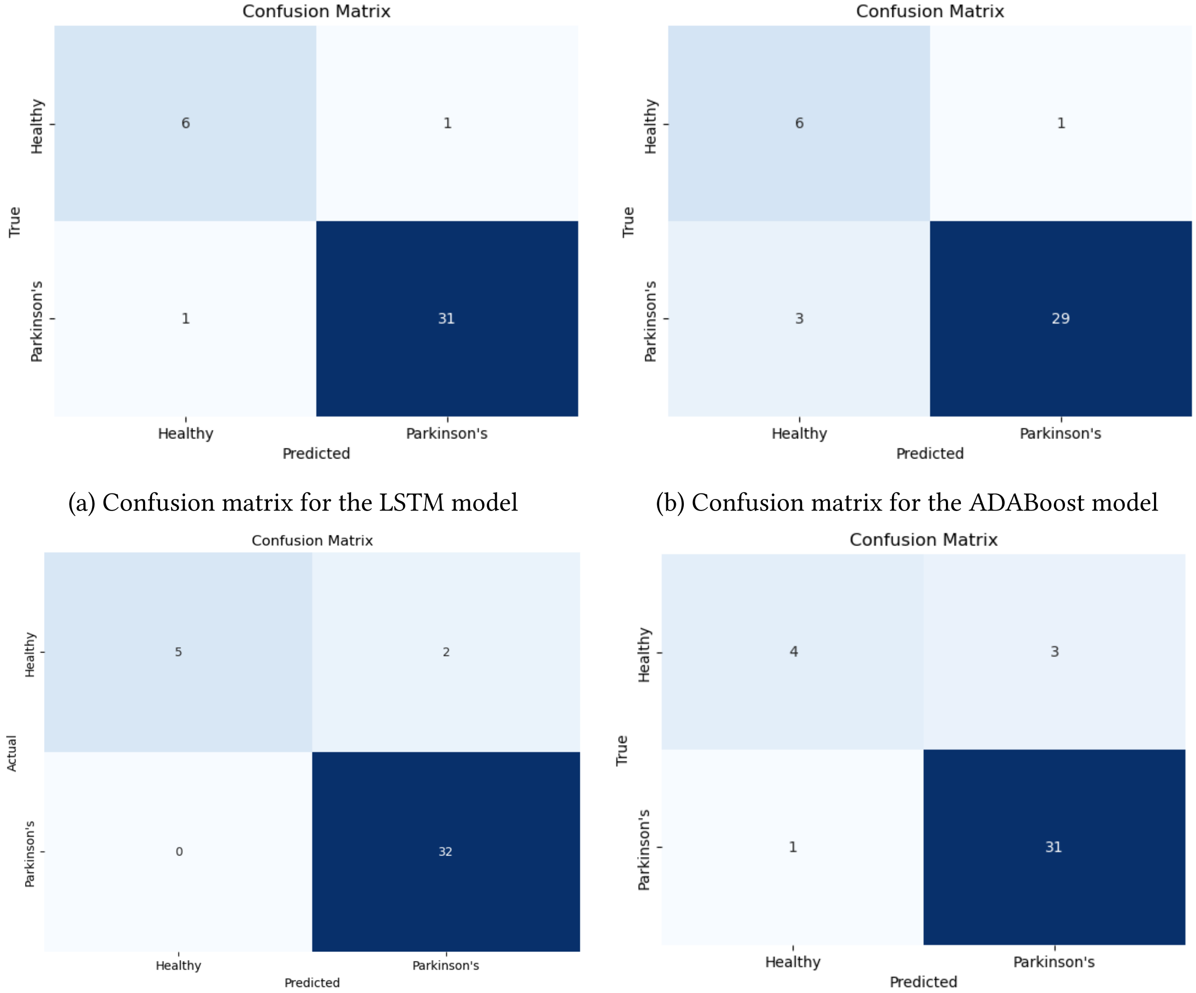
Long Short-Term Memory (LSTM) model emerges as the frontrunner, boasting an impressive accuracy of 0,97 and an F1 score of 0.98. This underscores the LSTM's prowess in effectively discerning positive instances of PD. With a precision of 0.96, the LSTM model showcases a remarkable ability to classify positive predictions accurately. It achieves a commendable recall rate of 0.96, indicating its ability to capture a high percentage of positive cases. The Random Forest and AdaBoost models exhibit lower accuracies of 0.89 and similar F1 scores of 0.93. In contrast, the Gradient Boosting model stands out with an accuracy of 0.94 and an impressive F1 score of 0.96, attributed to its remarkable precision and recall rates of 0.94 and 0.99, respectively. These results highlight the potential of the LSTM model as a significant contributor to PD detection, with potential implications for advancing early diagnosis and intervention strategies in clinical settings.

| Model | Accuracy | F1score | precision | Recall |
|---|---|---|---|---|
| LSTM model | 0.97 | 0.98 | 0.96 | 0.96 |
| R.forest | 0.89 | 0.93 | 0.91 | 0.96 |
| Gradient boosting | 0.94 | 0.96 | 0.94 | 0.99 |
| Adaboost | 0.89 | 0.93 | 0.96 | 0.90 |

**Table 1**
Comparison of various models

### 5.3. Confusion matrices

Confusion matrix is a popular tool for assessing how accurate and correct a model is, especially for classification tasks with many classes. Although the confusion matrix does not provide a complete performance statistic by itself, it is the basis of many performance measurements. In this work, the confusion matrix is illustrated in Figure 2; thus, the performance of the model can also conveniently be observed.



(a) Confusion matrix for the LSTM model

(b) Confusion matrix for the ADABoost model

(c) Confusion matrix for the Gradient boosting model

(d) Confusion matrix for the Random forest model

**Figure 2:** confusion matrix for each model.

## 6. Conclusion and future work:

Findings revealed that we were able to bring deep learning to the analysis of speech recordings, which enabled us to identify people with Parkinson's disease correctly. This work verifies the performance of the LSTM, Random Forest, Gradient Boosting, and AdaBoost models as used in the task of identifying Parkinson's disease through speech data. From the data, the LSTM model outperforms the others by a big margin in accuracy, precision, and recall. It shows that LSTM has much potential for early diagnosis of Parkinson's disease. In future studies, researchers should concentrate on optimizing the LSTM model and considering its potential clinical application for the application of earlier detection of Parkinson's disease. Moreover, our work shows why deep learning is essential for detecting Parkinson's disease.

## References

[1] G. Abdurrahman, M. Sintawati, Implementation of xgboost for classification of parkinson's disease, in: Journal of Physics: Conference Series, volume 1538, IOP Publishing, 2020, p. 012024.

[2] C. Perez, Y. Campos-Roca, L. Naranjo, J. Martin, Diagnosis and tracking of parkinson's disease by using automatically extracted acoustic features, J Alzheimers Dis Parkinsonism 6 (2016) 2161–0460.

[3] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E.

Isenkul, H. Apaydin, A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform, Applied Soft Computing 74 (2019) 255–263.

[4] T. Bocklet, E. Nöth, G. Stemmer, H. Ruzickova, J. Rusz, Detection of persons with parkinson's disease by acoustic, vocal, and prosodic analysis, in: 2011 IEEE Workshop on Automatic Speech Recognition Understanding, 4, 2011, pp. 478–483. doi:10.1109/ASRU.2011.6163978.

[5] R. Mathur, V. Pathak, D. Bandil, Parkinson disease prediction using machine learning algorithm, in: Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, 5, Springer, 2019, pp. 357–363.

[6] M. A. Little *, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, Suitability of dysphonia measurements for telemonitoring of parkinson's disease, IEEE Transactions on Biomedical Engineering 56 (2009) 1015–1022. doi:10.1109/TBME.2008.2005954.

# Enhancing Medical Image Analysis through Geometric and Photometric Transformations

Khadija Rais[1], Mohamed Amroune[1], Mohamed Yassine Haouam[1]
[1]Laboratory of mathematics, informatics and systems (LAMIS),
Echahid Cheikh Larbi Tebessi University,
Tebessa, 12002, Algeria.
E-mails: khadija.rais@univ-tebessa.dz; mohamed.amroune@univ-tebessa.dz;
mohaùed-yassine.haouam@univ-tebessa.dz;

***Abstract*—Medical image analysis suffers from a shortage of labeled data due to several challenges including patient privacy and lack of experts, although some AI models perform well only with large amounts of data and here, we will move to data augmentation where there is a solution to improve the performance of our models and increase the dataset size through traditional or advanced techniques.**

**In this paper, we evaluate the effectiveness of data augmentation techniques on two different medical image datasets. In the first step, we applied some transformation techniques to the skin cancer dataset containing benign and malignant classes and then trained the convolutional neural network (CNN) on the dataset before and after augmentation, which resulted in a significant improvement in the test accuracy from 90.74% to 96.88% and a decrease in the test loss from 0.7921 to 0.1468 after augmentation. In the second step, we enter Mixup technique by mixing two random images and their corresponding masks using the retina and blood vessel dataset, then we trained the U-net model and obtained the Dice coefficient which was improved from 0 before augmentation to 0.4163 after augmentation.**

**The result shows the effect of using data augmentation to increase the dataset size on the classification and segmentation performance.**

***Index Terms*—Data augmentation, Segmentation, Classification, Mixup, Transformation techniques**

## I. Introduction

Deep learning has changed medical image analysis by allowing them to create automatic systems with extraordinary accuracy in detecting, classifying, and segmenting various diseases. This propped the innovations for use in clinical diagnosis and decision-making. However, one of the biggest obstacles to applying deep learning to medical imaging is the relative lack of annotated datasets. Medical imaging datasets are typically small due to the challenges associated with data acquisition and labeling, which often require domain expertise and a fair amount of time. This confines the immense capacity for training such robust models that can generalize into unseen data.

Deep learning models succeed greatly when large and diverse datasets are available. Small datasets cause the model to fit well on training data but perform poorly on test data, losing its applicability in real-life scenarios. Data augmentation provides an efficient method of overcoming such cases by artificially increasing the size and variability of datasets [1]. Simple yet effective traditional augmentations such as rotation, flipping, scaling, cropping, and color variations are also used for diversifying training data. They help generalize the model by exposing the network to various transformations of the input data during training and thus increasing its robustness to changes in real-life applications. Mixup is a promising technique in which two random images and their corresponding masks are mixed to create new training samples [2]. This method is smooth and simple compared to deep learning techniques including GAN and VAE [3], resulting in better generalization and robustness, and its benefits have been demonstrated in many different domains, with little impact on medical image analysis, especially segmentation tasks[4].

Our main contribution is to examine the impact of data augmentation techniques on two distinct medical imaging analysis tasks, namely classification and segmentation. We first implement traditional augmentation techniques on a skin cancer dataset containing two classes: benign and malignant. The dataset was distributed to train a Convolutional Neural Network (CNN) to classify skin lesions, and the performance of the model was assessed before augmentations and after augmentations. Secondly, we performed image segmentation by introducing the mixup technique on a retina-blood-vessel dataset using a U-Net model. The need for precise delineation of blood vessels is the main challenge of the segmentation task, and the mixup method introduces novel training samples that help address this challenge.

The results demonstrate the critical role of using data augmentation to improve the performance of medical analysis for both classification and segmentation tasks, with traditional data augmentation techniques demonstrating their improvements using the cancer dataset that requires high accuracy and low loss. Moreover, the mixing technique allows U-Net to perform well and obtain meaningful images for the retinal and vascular dataset. The structure of this paper consists of Section 2 for related work, then methodology in Section 3, discussion and results in Section 4, and finally conclusion in Section 5.

## II. Related works

Image augmentation increases the dataset size and quality for training by adding more examples to network [5]. Augmentation data techniques are partitioned by researchers into traditional and deep learning-based, where traditional methods involve modifications implemented on data through transformations like flipping, cropping, and noise injection. On the other hand, advanced augmentation methods may incorporate generative models to create entirely new data instances as opposed to mere transformations [6].

In the light of increasing diabetic retinopathy (DR), the necessity of diagnostic system that can cope with time-expensive physician examination and latent lesions (photocoagulation) is addressed, the researchers in [7] The EyePACS dataset was addressed to study the key aspects of such an automated diagnostic solution with focus on data preprocessing, affine transformations and overfitting or class imbalance. Deep learning architecture, which encompasses seven pre-trained deep CADs were evaluated for DR classification. The most performing model of all the architectures trained on EfficientNetV2-M with a test accuracy of 97.65% shows efficacy in this study. Classification metrics like precision, recall, F1 Score, accuracy and loss were utilized to measure the performance. The researchers in another study [8] examined the effect of seven known image augmentation methods on CNN performance in binary classification problems in eleven medical datasets (mainly lung infections and cancer datasets; This included X-ray, USg, PET/CT as well as MRI images). Input augmentation: Vertical and horizontal reflections, fixed random rotations, translations, and crops were tried. All augmentation leads to the creation of one extra copy of every training image in our dataset, thus double dataset. No statistical significance was observed for both US and PET datasets. But Gaussian blur was the best augmentation technique for X-rays and MRI images. The authors in this paper address issues of scarce and imbalanced medical image datasets via Siamese neural networks with approaches like data balancing, weighted loss and constrained augmentation. We achieve up to 5.6% F1 score gain over CNNs on various datasets for COVID-19 diagnosis in the experiments using chest X-ray datasets. Safe augmentations (restricted shifts, scaling and rotations) are employed to keep essential information and increase the robustness of the model, while avoiding any changes that may impact diagnostics [9].

LCAMix is a new technique for data augmentation in medical image segmentation that considers mixing images and masks with contour-aware, superpixel-based techniques. To allow for increased spatial awareness, this will introduce two auxiliary tasks: the classification of local superpixels and the reconstruction of source images. LCAMix is model agnostic, straightforward to implement, and does not need external data; it has shown to outperform various datasets [10]. Deep learning techniques have succeeded in segmenting organ systems but do encounter challenges in lesion segmentation due to data deficiency, morphology diversity, and a lack of informative features. To tackle the above challenges, this paper presents Self-adaptive Data Augmentation (SelfMix), an innovative way towards better lesion segmentation via the self-adaptive fusion of lesion and non-lesion information. Modelled as a unique process unlike existing techniques like Mixup, CutMix, and CarveMix, it contains three major standpoints: (1) less distortion introduced since both tumor and non-tumor information are injected; (2) inclusion into the formula of fusion weights adapted to the lesion geometry and size; (3) brings into consideration non-tumor information. Through experiments on two public datasets, it was shown that SelfMix considerably outperform existing methods in lesion segmentation accuracy [11]. This study investigates whether the mixup data augmentation technique originally designed for classification tasks can enhance the performance of deep segmentation networks in medical imaging. The researchers compared results of U-Net trained on 100 3D T2-weighted MRI scans with and without mixup for prostate segmentation. Metrics such as the Dice similarity coefficient and mean surface distance were used to assess performance. This mixup statistically yielded improvements of up to 1.9% Dice and 10.9% surface distance reduction compared to normal training, with the suggestion that it could be useful in alleviating the issue of data deficit in medical image segmentation [2]. The lack of data to train complex models was addressed by developing a novel dataset extracted from chest CT slices, with the Mixup data augmentation method integrated into a semi-supervised learning frame called Mixup-Inf-Net. Since this procedure utilizes the minimum amount of annotated data and also takes advantage of unlabeled-DATAs over Mixup, the identification of COVID-19-infected areas is possible. Testing on SemiSeg datasets with 3D CT images showed that Mixup-Inf-Net surpasses most of the state-of-the-art segmentation models and is an enhancement on performance learning [12]. This study examines the effect of the mixup data augmentation technique on enhancing the segmentation performance of the U-Net model. The dataset of histopathological images was classified into three groups: (1) augmented with traditional methods (flipping, rotation) images, (2) those augmented only using the mixup method, and (3) images augmented using both techniques. The results indicate that combining mixup with traditional augmentation methods provides an increase in the average Dice coefficient of the model for artifact segmentation [13]. Zhang et al. describe CarveMix, a data augmentation technique used to facilitate convolutional neural network (CNN)-based brain lesion segmentation, which focuses on preserving lesion information through harmonization steps for heterogeneous data and models the mass effect unique to whole brain tumor segmentation. CarveMix stochastically mixes two annotated brain lesion images by carving a ROI based on lesion location and geometry and replacing corresponding voxels in the second image. The results on the multiple datasets indicate that CarveMix successfully enhances the segmentation accuracy of brain lesions [14]. A new augmentation method called TensorMixup uses the three-dimensional U-Net architecture for brain tumor segmentation. The procedure involves selecting patches of image MRI data from two patients using the same modality and mixing them with a tensor that is distributed from a sample of Beta distribution while creating a new image and its corresponding one-hot encoded label. The model trained on this augmented data achieves good segmentation performance, achieving high Dice scores of 92.15%, 86.71%, and 83.49% used for whole tumor, tumor core, and enhancing tumor segmentation, respectively, demonstrating the efficacy of TensorMixup [15].

## III. Methodology

### A. Data augmentation

Our work aims to improve medical image analysis through data augmentation. Geometric transformation including rotation, scaling, translation, shearing, reflection and optical transformations including brightness adjustment, contrast adjustment, Gaussian noise and histogram equalization are applied to a dataset of benign skin images to generate various variables, as shown in Figure 1: Samples of geometric and photographic transformation. The augmented images are stored for further training of the model.

Geometric Transformations: The following transformations were applied to each image:

- Rotation: Images rotated 90 degrees clockwise.
- Scaling: Images were resized with a scaling factor of 1.2 to simulate zoom effects.
- Translation: Images were shifted by 20 pixels horizontally and 30 pixels vertically.
- Shearing: A shear transformation matrix was applied to create skew effects.
- Flipping: Images were horizontally flipped to create mirrored versions.

Photometric Transformations: The following intensity-based transformations were performed to simulate variations in lighting and contrast:

- Brightness Adjustment: Brightness levels were increased using pixel value adjustments.
- Contrast Adjustment: Contrast levels were amplified using scaling factors.
- Gaussian Noise Addition: Random Gaussian noise was added to simulate imaging artifacts.
- Histogram Equalization: The luminance channel of the images was equalized to enhance contrast while preserving color fidelity.

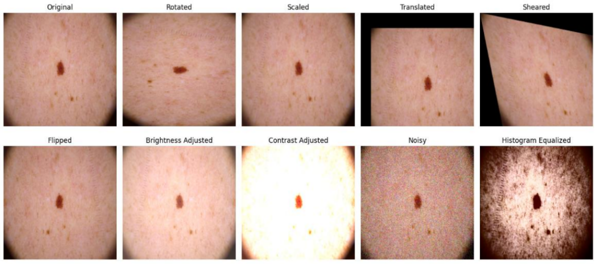


Fig. 1. Samples of geometric and photographic transformation

Mixup: To simulate realistic lesions and interactions between images, a Mixup strategy was implemented. This method combined two images and their corresponding masks to generate new synthetic samples, as shown in Figure 2, preserving lesion information.

Two randomly selected images, along with their masks, were processed:

- Lesion areas were extracted from the first image using the corresponding binary mask.
- Background areas were extracted from the second image.
- A Mixup coefficient ($\lambda$) was sampled from a Beta distribution ($\alpha$=0.4) controlling the blend ratio.
- A weighted combination of lesion and background areas was computed for both images and masks, creating mixed images and masks.

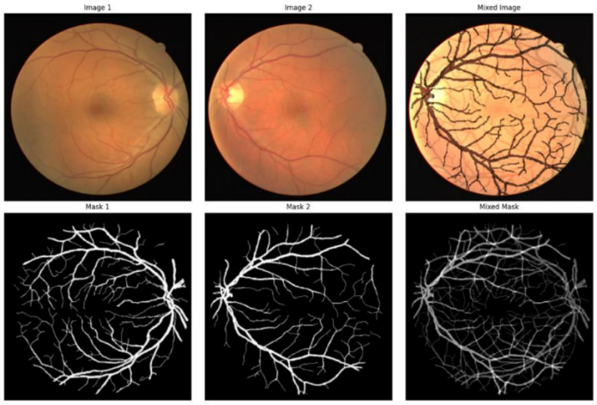


Fig. 2. Mixup of two images

### B. Dataset

In this study, we used two Kaggle datasets. The first dataset contains skin cancer images [16], which were curated to support tasks such as image classification. They were classified as malignant using 240 images and benign using 30 images. This class was expanded to achieve balance, resulting in 240 images to support accurate discrimination between harmful and harmless parasites. The second dataset shows blood vessels [17] in retinal fundus images, focuses on the segmentation task, contains 100 images and 100 masks, and is divided into 80% images for training and 20% for testing, including their corresponding masks. To enrich this training set, an additional 100 images were augmented.

## IV. Results and discussion

The results for the skin lesion classification task demonstrate the significant impact of data augmentation on CNN performance. Before applying data augmentation, the model achieved an accuracy of 90.74% with a loss of 0.7921. This performance was limited by the severe class imbalance, with only 30 benign images compared to 240 malignant images.

After balancing the dataset through data augmentation, the model's accuracy improved to 96.88%, and the loss decreased to 0.1468, indicating better generalization and robustness. This underscores the importance of data augmentation in handling imbalanced datasets and enhancing the reliability of medical image classification models, Figure 5 presents the losses and accuracies for test data before adding augmented data into train data and Figure 6 presents the losses and accuracies for test data after data augmentation.

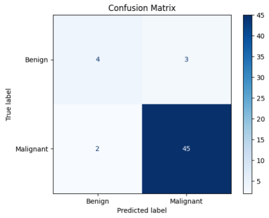


Fig. 3. Confusion matrix before data augmentation

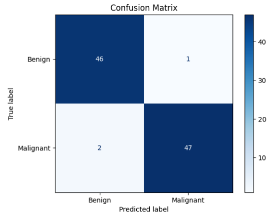


Fig. 4. Confusion matrix with data augmentation

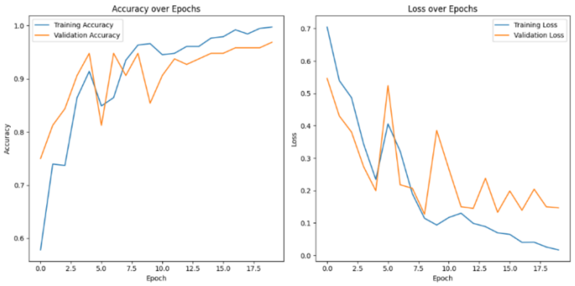


Fig. 5. Training history before data augmentation

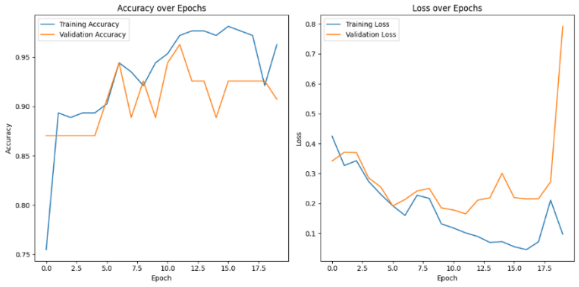


Fig. 6. Training history with data augmentation

For the segmentation task on retinal blood vessels, the dataset initially contained 80 training images and 20 test images with corresponding masks. Before applying data augmentation, the model failed to achieve meaningful segmentation, resulting in a Dice coefficient of 0.00.

By augmenting the training set to include 180 images, the Dice coefficient improved to 0.4163, as mentioned in Table 1, showing a marked improvement in segmentation performance. The additional data likely helped the model learn more diverse patterns, leading to better vessel detection, Figure 7 presented some predicted masks and the dice range from 0.3370 to 0.5887.

The results highlight the critical role of data augmentation in medical image analysis, particularly for datasets with limited or imbalanced samples. For classification, data augmentation not only corrected the imbalance but also enhanced the model's ability to distinguish between benign and malignant lesions. Similarly, in segmentation, augmenting the dataset introduced variability that helped the model achieve more accurate blood vessel segmentation.

However, despite improvements, the Dice score for segmentation remains modest, suggesting room for further enhancement. Techniques like advanced augmentation strategies, better model architectures, or fine-tuning hyperparameters could be explored to improve performance further.

These findings emphasize the necessity of data augmentation in medical image tasks to improve model accuracy and reliability, ultimately aiding in early diagnosis and treatment planning.

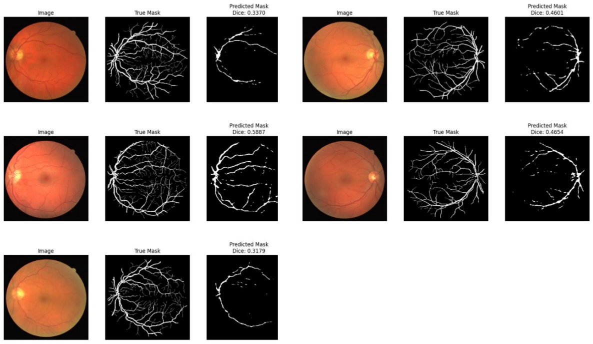


Fig. 7. Predicted masks

TABLE I
CLASSIFICATION AND SEGMENTATION TASK WITH AND WITHOUT DATA AUGMENTATION

| Task | Data augmentation (DA) | Dataset | No. of images | Metrics |
|---|---|---|---|---|
| Classification with CNN | Before DA | Skin cancer | Benign: 30 | Accuracy: 90.74%<br>Loss: 0.7921 |
| | | | Malignant: 240 | |
| | With DA | Skin cancer | Benign: 240 | Accuracy: 96.88%<br>Loss: 0.1468 |
| | | | Malignant: 240 | |
| Segmentation with U-Net | Before DA | Blood vessels in retinal | Train:<br>Mask Images: 80 80<br>Test:<br>Mask Images: 20 20 | Mean Dice: 0.00 |
| | With DA | Blood vessels in retinal | Train:<br>Mask Images: 180 180<br>Test:<br>Mask Images: 20 20 | Mean Dice: 0.4163 |

## V. CONCLUSION

This work underlines the essential role for data augmentation in improving medical image analysis performance, when it is limited or imbalanced datasets that we are dealing with. In the case of skin cancer classification, data augmentation successfully dealt with class imbalance that achieved an accuracy of 96.88% up from 90.74% and significantly reduced the loss.

For Retinal Blood Vessel Segmentation, data augmentation on the dataset size and variability lifted making Step from 0.00 to a notable Dice coefficient win of 0.4163.

These results indicate data augmentation can potentially help to ease model overfitting, providing more robust and accurate results to difficult medical image tasks. Ideally, we wish to consider future directions with more elaborate augmentations and architecture, such as performance enhancement or dataset diversity. The current study once again confirms the significance of data augmentation as a pre-processing step to generate robust AI models for medical image analysis and contribute early detection and improved diagnostic outcomes.

## Bibliography

## References

[1] K. Rais, M. Amroune, and M. Y. Haouam, ‘Medical Image Generation Techniques for Data Augmentation: Disc-VAE versus GAN’, in 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), 2024, pp. 1–8. doi: 10.1109/PAIS62114.2024.10541221.

[2] L. J. Isaksson et al., ‘Mixup (sample pairing) can improve the performance of deep segmentation networks’, J. Artif. Intell. Soft Comput. Res., vol. 12, no. 1, pp. 29–39, 2022, doi: 10.2478/jaiscr-2022-0003.

[3] K. Rais, M. Amroune, A. Benmachiche, and M. Y. Haouam, ‘Exploring Variational Autoencoders for Medical Image Generation: A Comprehensive Study’, ArXiv Prepr. ArXiv241107348, 2024, doi: 10.48550/arXiv.2411.07348.

[4] K. Rais, M. Amroune, M. Y. Haouam, and I. Bendib, ‘Comparative Study of Data Augmentation Approaches for Improving Medical Image Classification’, in 2023 International Conference on Computational Science and Computational Intelligence (CSCI), 2023, pp. 1226–1234. doi: 10.1109/CSCI62032.2023.00200.

[5] Z. Yang, R. O. Sinnott, J. Bailey, and Q. Ke, ‘A survey of automated data augmentation algorithms for deep learning-based image classification tasks’, Knowl. Inf. Syst., vol. 65, no. 7, pp. 2805–2861, 2023, doi: 10.1007/s10115-023-01853-2.

[6] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, and M. Mridha, ‘A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions’, Healthc. Anal., p. 100340, 2024, doi: 10.1016/j.health.2024.100340.

[7] R. İncir and F. Bozkurt, ‘A study on effective data preprocessing and augmentation method in diabetic retinopathy classification using pretrained deep learning approaches’, Multimed. Tools Appl., vol. 83, no. 4, pp. 12185–12208, 2024, doi: 10.1007/s11042-023-15754-7.

[8] O. Rainio and R. Klén, ‘Comparison of simple augmentation transformations for a convolutional neural network classifying medical images’, Signal Image Video Process., vol. 18, no. 4, pp. 3353–3360, 2024, doi: 10.1007/s11760-024-02998-5.

[9] A. Galán-Cuenca, A. J. Gallego, M. Saval-Calvo, and A. Pertusa, ‘Few-shot learning for COVID-19 chest X-ray classification with imbalanced data: an inter vs. intra domain study’, Pattern Anal. Appl., vol. 27, no. 3, p. 69, 2024, doi: 10.1007/s10044-024-01285-w Download citation.

[10] D. Sun, F. Dornaika, and J. Charafeddine, ‘LCAMix: Local-and-contour aware grid mixing based data augmentation for medical image segmentation’, Inf. Fusion, vol. 110, p. 102484, 2024, doi: 10.1016/j.inffus.2024.102484.

[11] Q. Zhu, Y. Wang, L. Yin, J. Yang, F. Liao, and S. Li, ‘Selfmix: a self-adaptive data augmentation method for lesion segmentation’, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 683–692. doi: 10.1007/978-3-031-16440-8_65.

[12] K. Zeng, X. Pei, and P. Chen, ‘Mixup-Inf-Net: a data augmentation algorithm for segmentation of new coronary pneumonia infections’, in International Conference on Signal Processing, Computer Networks, and Communications (SPCNC 2022), SPIE, 2023, pp. 132–137. doi: 10.1117/12.2674401.

[13] S. A. Arpaci and S. Varli, ‘Semantic segmentation with the mixup data augmentation method’, in 2022 30th Signal Processing and Communications Applications Conference (SIU), IEEE, 2022, pp. 1–4. doi: 10.1109/SIU55565.2022.9864873.

[14] X. Zhang et al., ‘CarveMix: a simple data augmentation method for brain lesion segmentation’, NeuroImage, vol. 271, p. 120041, 2023, doi: 10.1016/j.neuroimage.2023.120041.

[15] Y. Wang, Y. Ji, and H. Xiao, ‘A data augmentation method for fully automatic brain tumor segmentation’, Comput. Biol. Med., vol. 149, p. 106039, 2022, doi: 10.1016/j.compbiomed.2022.106039.

[16] ‘Skin Cancer Dataset’.

[17] ‘Retina Blood Vessel’.

# Ensuring Data Security and Privacy in Smart City Surveillance Systems Using Chaotic Maps and Smart Contracts⋆

Mohamed ElAmine Kheraifia, Abdelatif Sahraoui , Makhlouf Dardour , Abdelghani Guerrad and Ahmed Arbaoui

[1] *Cheikh Larbi Tebessi University LAMIS Laboratory,Tebessa, 12000 , Algeria*
[2] *Cheikh Larbi Tebessi University LAMIS Laboratory,Tebessa, 12000 , Algeria*
[3] *University Of Oum el Bouaghi LIAOA Laboratory, Oum el Bouaghi , 04000 , Algeria*
[4] *Cheikh Larbi Tebessi University LAMIS Laboratory,Tebessa, 12000 , Algeria*
[5] *Cheikh Larbi Tebessi University LAMIS Laboratory,Tebessa, 12000 , Algeria*

**Abstract**

Smart cities leverage technology to enhance quality of life, strengthen security, and optimize resource management. In this context, video surveillance systems are essential, providing real-time monitoring, data collection, and intelligent insights to support various urban operations. However, despite their advantages, these systems also pose challenges, particularly related to privacy, security, and ethical concerns. This paper introduces a lightweight encryption approach that combines blockchain smart contracts with chaotic maps to effectively protect video records.

**Keywords**

Surveillance system, Chaotic maps, Smart contract

## 1. Introduction

A surveillance system in a smart city is an advanced network of cameras, sensors, and data-processing technologies aimed at improving security, public safety, and urban management. These systems leverage real-time video monitoring, data analytics, and artificial intelligence to detect anomalies, manage traffic, oversee public spaces, and respond to emergencies. When integrated with other smart city infrastructure, they also contribute to resource optimization, public health initiatives, and environmental monitoring. However, these systems pose significant risks, particularly regarding privacy and data security. The large volumes of sensitive information they collect make them attractive targets for cyberattacks. Without robust security measures, unauthorized access can lead to data breaches, jeopardizing citizens' privacy.Encryption plays a key role in ensuring the confidentiality, integrity, and security of this data by preventing malicious actors from accessing or manipulating the footage. However, traditional encryption methods come with challenges,This includes high computational costs, as traditional encryption algorithms like AES or RSA can demand substantial processing power,

*Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY*

⋆ You should use this document as the template for preparing your publication. We recommend using the latest version of the CEURART style.

* Corresponding author.

† These authors contributed equally.

mohammedelamine.kheraifia@univ-tebessa.dz (M.E.A. Kherifia); abdelatif.sahraoui@univ-tebessa.dz (A.Sahraoui); derdour.makhlouf@univ-oeb.dz (M. Derdour) ; abdelghani.guerrad@univ-tebessa.dz(A. Guerrad), ahmed.arbaoui@univ-tebessa.dz

0009-0002-2974-804X (M.E.A. Kherifia) ; 0000-0001-7254-6828 (A.Sahraoui); 0000-0001-6622-4355 (M. Derdour); 0009-0006-1776-0622(A. Guerrad) ; 0009-0003-3359-6031(A. Arbaoui)

particularly for high-definition video streams. This added load increases system latency, potentially affecting real-time monitoring. Additionally, key management becomes more complex with conventional encryption methods, as securely generating, distributing, and maintaining keys is challenging, especially in distributed surveillance networks with numerous cameras.

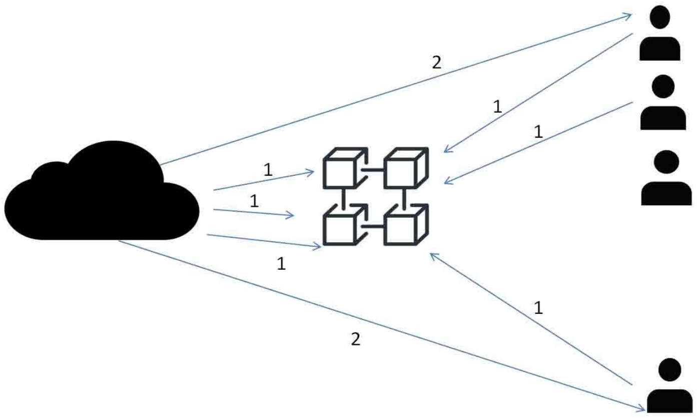


**Figure1**:Smart contract for generating initial chaotic map values between users and the cloud

We propose a lightweight system that combines chaotic maps with smart contracts. Chaotic maps, known for their sensitivity to initial conditions and unpredictable behavior, offer an efficient alternative for cryptographic applications. Unlike traditional methods, they enable high-speed key generation, minimizing computational overhead. Their efficiency makes them ideal for real-time encryption of video frames, even in resource-constrained environments. The encryption key for each video frame is determined by specific initial values, and slight variations in these values produce entirely different encryption outputs, ensuring robust security. Effective key management is crucial in any cryptographic system. In our approach, blockchain smart contracts manage and assign the initial values for chaotic map functions figure1 ensuring secure and decentralized control of the encryption process.

The content of this paper is organized as follows: Section II introduces the related work. Section III introduces the video surveillance system and chaotic maps. Section IV introduces our encryption system for video surveillance. Section V presents the performance evaluation of the proposal. Section VI concludes our work.

## 2. Related Work

Cryptography significantly enhances the security of online video streaming services by employing various advanced techniques to protect content from unauthorized access and piracy. Key methodologies include multi-key hybrid cryptography, which utilizes dynamic keys for real-time encryption of video chunks, thereby improving security and performance [1], Additionally, the AES-Rijndael algorithm is effective in safeguarding video data during transmission, ensuring that only authorized users can decrypt and view the content[2]. Furthermore, the integration of steganography within video streams adds an extra layer of security by embedding secret data within the video itself [3]. Lastly, a combination of symmetric and asymmetric cryptographic algorithms can optimize both security and processing speed, making it suitable for high-bandwidth applications like video streaming[4].While these cryptographic methods enhance security, they also introduce complexities in key management and may impact performance if not implemented efficiently. Balancing security with user experience remains a critical challenge in the development of streaming services.

Blockchain technology can indeed be utilized to create a secure and decentralized video sharing platform. This approach leverages a peer-to-peer (P2P) distribution model, eliminating the need for central servers and intermediaries, thus enhancing user control over content.Blockchain's cryptographic functions, such as hashing and signing, provide tamper-proof solutions, enhancing security and privacy for users[5].

Ghimire et al[6].,introduce a new video integrity verification (IVM) method that utilizes the blockchain framework. This approach employs an efficient blockchain model for centralized video data, combining a hash-based message authentication code with elliptic curve cryptography to ensure video integrity,[7]We proposed a video fingerprinting method that uses timestamps to prevent and detect image tampering or the substitution of original images during the transmission and reception of monitored data. This approach leverages blockchain technology for tracking, verification, and authentication processes, Jeong wt al[8]., The proposed system comprises a blockchain network managed by trusted internal authorities. Video metadata is securely recorded on the blockchain's distributed ledger, preventing data tampering. The architecture encrypts and stores the video, establishes authorization within the blockchain, and facilitates video export, Fitwi et al[9]., proposed a mechanism for securely sharing privacy-preserving surveillance videos by utilizing blockchain, smart contracts, and encryption techniques such as the Discrete Cosine Transform (DCT), Advanced Encryption Standard (AES), and a block shuffling algorithm for video frames, Gallo et al[10]., We present BlockSee, a blockchain-based video surveillance system that simultaneously verifies the authenticity and stability of both camera settings and surveillance footage, ensuring secure access for authorized users during critical events , Dhar et al[11]., presented a study on enhancing the security of multimedia data, including audio, video, and images, derived from IoT devices. Advanced technologies such as blockchain and quantum cryptography are being explored as promising solutions to enhance multimedia security and protect privacy, Kumar et al[12]., proposed a decentralized peer-to-peer platform for photo and video sharing, leveraging the Interplanetary File System (IPFS) and built on blockchain technology. The platform employs cognitive hashing (pHash) to detect and prevent multimedia copyright infringement.

## 3. Background

This section explores the video surveillance system, its components, and various techniques utilized in chaotic maps.

### 3.1. Video Surveillance System

A smart city video surveillance system is a sophisticated network of cameras and sensors aimed at monitoring public spaces and urban infrastructure. These systems utilize real-time monitoring, data analytics, and, in some cases, artificial intelligence to improve security, safety, and overall urban management. Beyond security, the data collected is leveraged for traffic management, environmental monitoring, public health, and emergency response efforts. In smart cities, the surveillance infrastructure is organized into four layers [13]. The primary layer includes surveillance equipment such as closed-circuit television (CCTV) systems, which transmit video signals to a limited number of devices. This layer incorporates various camera types, including IP cameras (Internet Protocol cameras) that transmit and receive data over networks. PTZ (Pan-Tilt-Zoom) cameras provide advanced functionality, allowing for horizontal (pan), vertical (tilt) movement, and zoom, making them ideal for monitoring expansive areas. Additionally, portable devices like dash cams, drones, and smartphones are part of this layer. The second layer, Edge Computing, manages the processing and analysis of video data close to its source, enabling real-time data processing. The third layer incorporates cloud computing, offering a scalable, flexible, and cost-efficient solution for storing video recordings. It enables convenient access to footage from anywhere with an Internet connection, The final layer, Blockchain, serves as a decentralized, secure, and immutable ledger. It is employed to prevent data tampering and unauthorized access, making it an ideal solution for video surveillance systems in smart cities. Permitted or private blockchains are often preferred due to their key features, including: Authorization and Authentication, Data integrity, Distributed storage and Smart Contracts.

### 3.2. Chaotic maps

Chaotic maps in cryptography are mathematical functions that strengthen encryption algorithms by introducing nonlinearity, sensitivity to initial conditions, and unpredictability—key attributes of chaos theory. Although chaotic systems are deterministic, they exhibit random-like behavior, making them ideal for cryptographic tasks such as random sequence generation, secure key creation, and encryption:

- **Logistic Map**: One of the simplest chaotic maps, defined by the equation:

$$x_{n+1} = rx_n(1 - x_n) \qquad (1)$$

where $r$ is a control parameter. The logistic map is widely used for generating chaotic sequences that serve as encryption keys or masking data.

- **Henon Map**: A two-dimensional chaotic map given by the equations[14]

$$\begin{gathered} x_{n+1} = 1 - ax_n^2 + y_n \\ y_{n+1} = bx_n \end{gathered} \qquad (2)$$

where $a$ and $b$ are control parameters. It is commonly used in image encryption due to its ability to generate complex chaotic behavior.

- **Arnold's Cat Map:** A chaotic map that scrambles images and can be reversed to restore the original image. It’s particularly used in image encryption[15]:

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix} (mod\ N) \qquad (3)$$

## 4. COMBINATION BETWEEN BLOCKCHAIN AND CHAOTIC MAPS

This section introduces a model we developed to safeguard video data from tampering. Encryption is essential for maintaining the integrity and reliability of surveillance systems, particularly in settings where privacy and data protection are paramount.

Figure 2 illustrates the steps of the proposed approach, where blockchain technology is combined with chaotic maps to encrypt video clips for a smart city surveillance system. The video clips are securely stored in the cloud. Chaotic maps generate sequences that appear random and are highly unpredictable, making them valuable for cryptographic applications like secure pseudo-random number generation.

Their sensitivity to initial conditions figure 3. makes chaotic maps ideal for producing encryption keys that are challenging to forecast. We initiated the process by developing a smart contract between the user and the cloud to generate variables for the chaotic map function and record their execution times . In the second stage, both the user and the cloud begin generating keys based on the settings and variables created by the smart contract, utilizing one of the chaotic map methods In the third and fourth steps, the cloud encrypts the video frames and sends them to the user. In the fifth step, the user decrypts the frames using the key, calculates the hash, and compares it with the timestamp file on the blockchain to verify their authenticity.

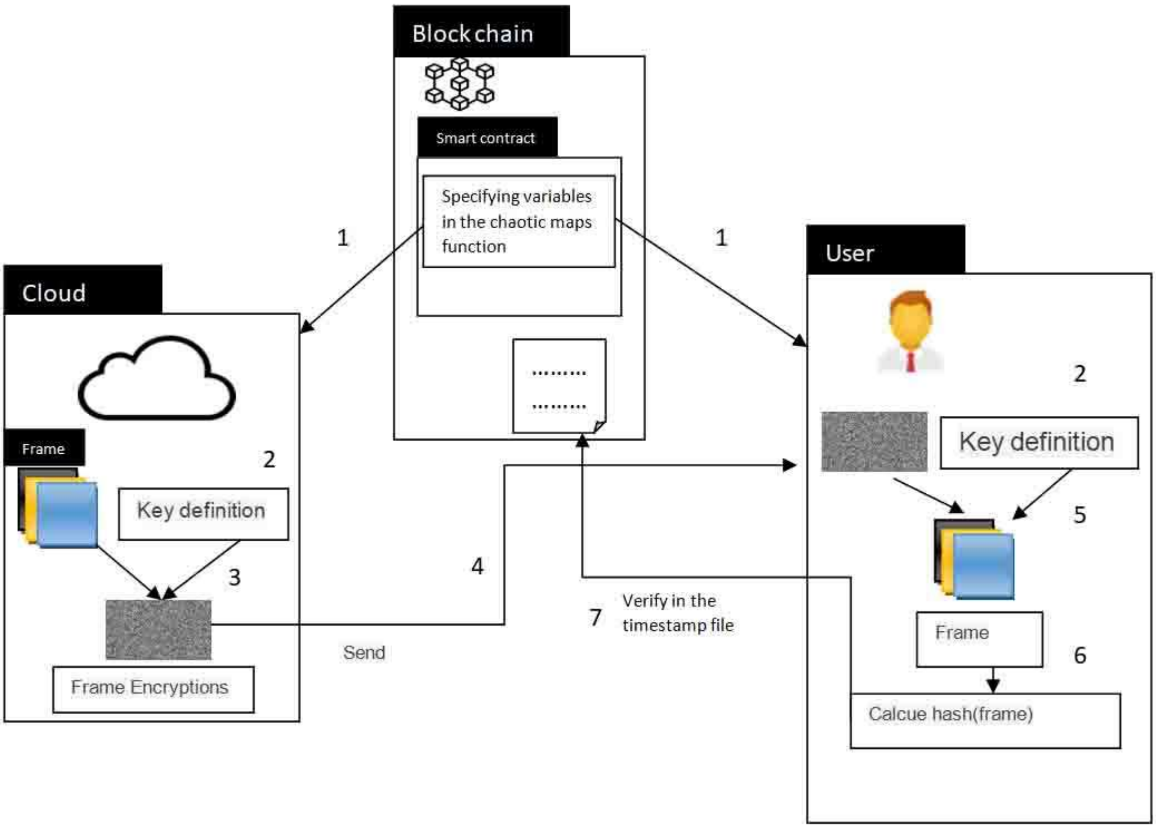
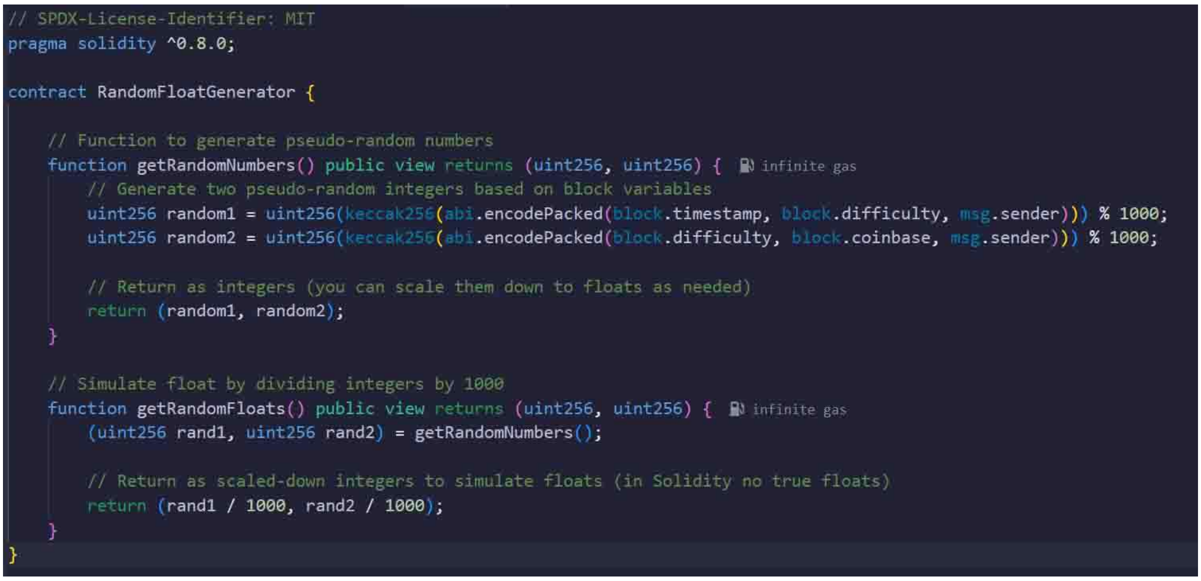


**Figure2**: Combination between Blockchain and Chaotic maps



```
// SPDX-License-Identifier: MIT
pragma solidity ^0.8.0;

contract RandomFloatGenerator {

    // Function to generate pseudo-random numbers
    function getRandomNumbers() public view returns (uint256, uint256) {  infinite gas
        // Generate two pseudo-random integers based on block variables
        uint256 random1 = uint256(keccak256(abi.encodePacked(block.timestamp, block.difficulty, msg.sender))) % 1000;
        uint256 random2 = uint256(keccak256(abi.encodePacked(block.difficulty, block.coinbase, msg.sender))) % 1000;

        // Return as integers (you can scale them down to floats as needed)
        return (random1, random2);
    }

    // Simulate float by dividing integers by 1000
    function getRandomFloats() public view returns (uint256, uint256) {  infinite gas
        (uint256 rand1, uint256 rand2) = getRandomNumbers();

        // Return as scaled-down integers to simulate floats (in Solidity no true floats)
        return (rand1 / 1000, rand2 / 1000);
    }
}
```

**Figure3**: The smart contracts in Solidity

### 4.1. Encrypted and Decrypted video frame

The figure 4 illustrates the method used to encrypt video frames. A key is generated from the initial values provided by the smart contract, ensuring that the key matches the frame size, forming a three-layer matrix. The encryption process involves combining the frame matrix with the key matrix in each layer. The remainder of division by 255 is then applied to produce the encrypted frame. The decryption process is the reverse of encryption. Upon receiving the encrypted frame, a key is generated using the initial values from the smart contract, ensuring the key is the same size as the encrypted frame. The decryption involves combining the encrypted frame matrix with the key matrix in each layer. The remainder of division by 255 is then applied to retrieve the original frame.

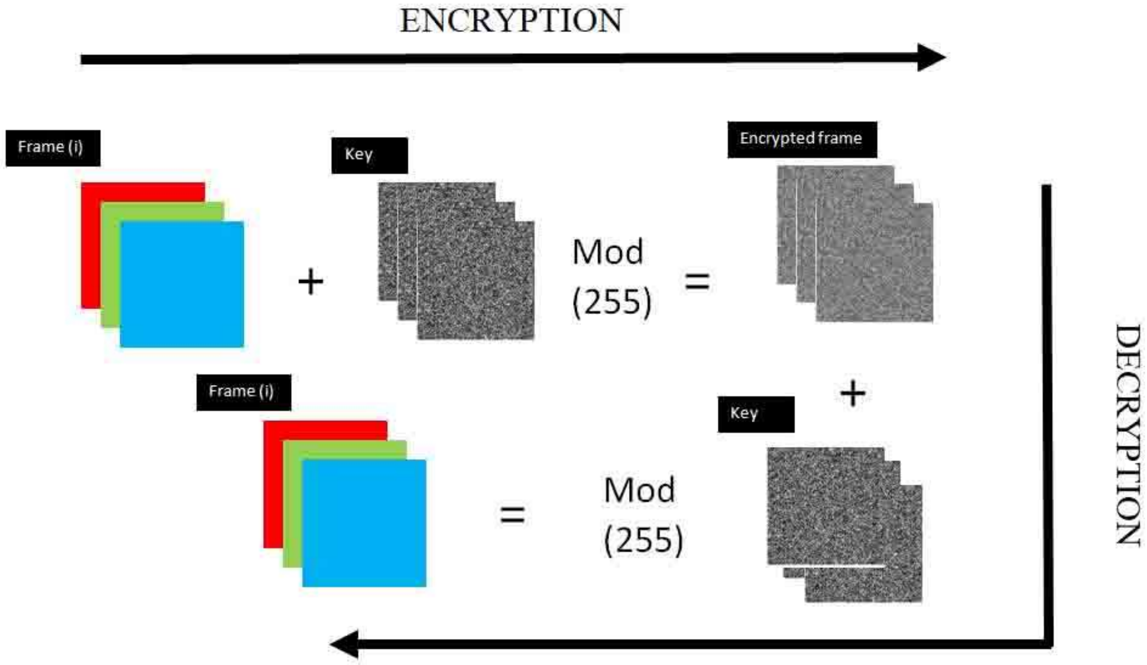


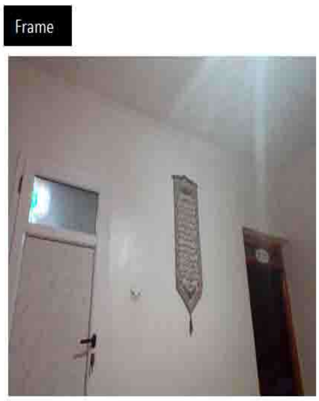
**Figure4**: Encrypted and Decrypted video frame

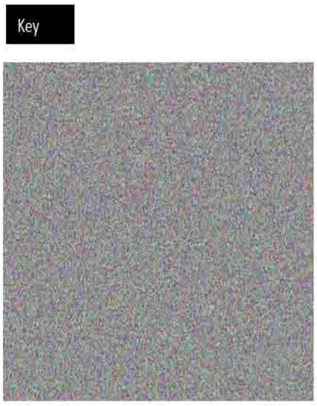
## 5. Evaluation and results

This section presents the results achieved. In our experience, we utilized the Ethereum blockchain, a decentralized platform that enables the development and implementation of security solutions for video surveillance through smart contracts involving blockchain users, IoT devices, edge nodes, and cloud storage services. These smart contracts were developed in Solidity, a language specifically designed for creating smart contracts on blockchain platforms like Ethereum. Additionally, we employed the C# programming language, developed by Microsoft. This modern, flexible, and object-oriented language was used to formulate security policies suitable for surveillance, including applications for mobile and cloud environments. The table 1 shows a comparison between different calculations in resource usage: amount of RAM used, percentage of CPU usage, and average time per iteration over 1000 iterations, based on the algorithm used to calculate the AES , and CHAOTIC MAP

| | **AES** | **CHAOTIC MAP** |
|---|---|---|
| **CPU USAGE(%)** | 9% | 12% |
| **MEMORY USAGE(MB)** | 7$(mb)$ | 4$(mb)$ |
| **AVERDGE TIME PER ITERATION** | 0.116525ms | 0.0012003ms |

**TABLE1**: Resource Usage Comparison





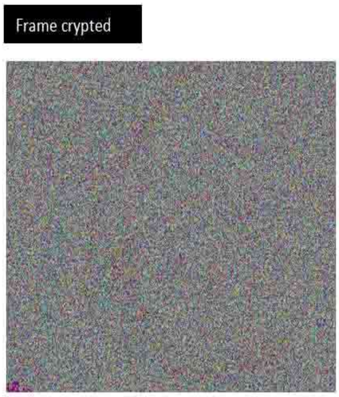


**Figure5**: Origin frame

**Figure6**: The key

**Figure7**: The encrypted frame

The figures illustrate the results obtained. The figure 5 displays the original video frame, the The figure 6 shows the key used for encryption, and the figure 7presents the resulting encrypted frame

## 6. Conclusion

The integration of chaotic maps and blockchain smart contracts offers a lightweight, secure, and efficient solution for encrypting video in surveillance systems. This approach addresses the challenges of traditional encryption methods by reducing processing demands and streamlining key management. By leveraging blockchain's decentralized and tamper-proof nature, the system ensures the integrity of video data, providing a reliable framework for safeguarding surveillance footage in smart city environments. This combination of technologies ensures that sensitive data remains protected from unauthorized access, tampering, and breaches, promoting privacy and security in the digital age.

## References

[1] Fouzar, Y., Lakhssassi, A., & Ramakrishna, M. (2023). A novel hybrid multikey cryptography technique for video communication. IEEE Access,11, 15693-15700.

[2] Zebedeus, Cheyso., I, Made, Agus, Dwi, Suarjaya., Gusti, Made, Arya, Sasmita. (2022). 3. Video Streaming Data Security Using AES-Rijndael Algorithm. CESS (Journal of Computer Engineering, System and Science),doi: 10.24114/cess.v7i2.32989

[3] Joseph, Scott, Morton., Christopher, Michael, McDonald., Glenn, Donald,Knepp. (2015). 4. Video cryptography system and method.

[4] Toan, Van, Nguyen., Dang, Quoc, Minh, Do., Phuc, Duc, Nguyen.,Thuan, Huu, Huynh., Thuc, Dinh, Nguyen. (2015). 5. Designing a high performance cryptosystem for video streaming application. doi:10.32508/STDJ.V18I3.836

[5] Moolikagedara, K., Nguyen, M., Yan, W. Q., Li, X. J. (2023). Video Blockchain: A decentralized approach for secure and sustainable networks with distributed video footage from vehicle-mounted cameras in smart cities.Electronics, 12(17), 3621.

[6] Ghimire, S., Choi, J. Y., Lee, B. (2019). Using blockchain for improved video integrity verification. IEEE Transactions on Multimedia, 22(1), 108-121.

[7] Kheraifia, M. E. A., Sahraoui, A., & Derdour, M. (2024, April). Blockchain-Driven Adaptive Streaming for IoT: Redefining Security in Video Delivery.In 2024 6th International Conference on Pattern Analysis and IntelligentSystems (PAIS) (pp. 1-7). IEEE.

[8] Jeong, Y., Hwang, D., Kim, K. H. (2019, January). Blockchain-based management of video surveillance systems. In 2019 International Conference on Information Networking (ICOIN) (pp. 465-468). IEEE. K. Elissa, "Title of paper if known," unpublished

[9] Fitwi, A., Chen, Y. (2021, July). Secure and privacy-preserving stored surveillance video sharing atop permissioned blockchain. In 2021 International Conference on Computer Communications and Networks (ICCCN)(pp. 1-8). IEEE.

[10] Gallo, P., Pongnumkul, S., Nguyen, U. Q. (2018, June). BlockSee:Blockchain for IoT video surveillance in smart cities. In 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/ICPSEurope) (pp. 1-6). IEEE.

[11] Dhar, S., Khare, A., Dwivedi, A. D., Singh, R. (2024). Securing IoT devices:A novel approach using blockchain and quantum cryptography. Internet of Things, 25, 101019.

[12] Kumar, R., Tripathi, R., Marchang, N., Srivastava, G., Gadekallu, T. R., Xiong, N. N. (2021). A secured distributed detection system based on IPFS and blockchain for industrial image and video data security. Journal of Parallel and Distributed Computing, 152, 128-143.

[13] Ghimire, S., Choi, J. Y., Lee, B. (2019). Using blockchain for improved video integrity verification. IEEE Transactions on Multimedia, 22(1), 108-121.

[14] Wei-Bin, C., Xin, Z. (2009, April). Image encryption algorithm based onHenon chaotic system. In 2009 International Conference on Image Analysis and Signal Processing (pp. 94-97). IEEE

[15] Kocarev, L., Lian, S. (Eds.). (2011). Chaos-based cryptography: Theory, algorithms and applications (Vol. 354). Springer Science Business Media

# Exploring the Integration of Differential Privacy in Cybersecurity Analytics: Balancing Data Utility and Privacy in Threat Intelligence

Brahim Khalil Sedraoui[1], Abdelmadjid Benmachiche[2], Amina Makhlouf[3], and Chaouki Chemam[4]

[1] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*
[2] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*
[3] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*
[4] *University of Chadli Bendjedid, Faculty of Sciences & Technology, El Tarf, Algeria*

**Abstract**

In order to tackle the crucial problem of protecting privacy while guaranteeing data usefulness in threat intelligence, this study investigates the incorporation of Differential Privacy (DP) into cybersecurity analytics. A strong mathematical framework called Differential Privacy ensures privacy by introducing calibrated noise into data outputs, which stops sensitive information from leaking even when auxiliary datasets are present. The use of DP in Security Information and Event Management (SIEM) systems is highlighted in the article, which also demonstrates how DP may protect event log and threat data analysis without sacrificing analytical efficacy. The privacy-utility trade-offs involved in maximizing the epsilon parameter—a crucial part of DP mechanisms—are highlighted. through real-world applications and case studies. This article demonstrates the revolutionary potential of DP in promoting safe data sharing and cooperative threat intelligence through case studies and real-world applications. In the end, this study promotes Differential Privacy as a crucial tactic for improving privacy-preserving analytics in the field of cybersecurity.

**Keywords**

Differential Privacy, Cybersecurity Analytics, Privacy-Preserving Techniques.

## 1. Introduction to Differential Privacy and Cybersecurity Analytics

Differential privacy (DP) has become a widely accepted framework for preserving the privacy of individual entries in a dataset when statistical analyses of the data are published. DP was introduced in the computer science and statistics literature in 2006 for providing formal guarantees about the privacy of the individuals described by a dataset, while allowing for useful generalizations to be made and shared about that dataset [1].

As a generalization, the privacy of the individuals in a dataset is protected by noise addition. The fundamental DP framework describes a general process for sharing and analyzing a dataset were given a dataset $D$, a function $f$, and an ε > 0, for all neighboring datasets $D$ and $D0$ (i.e., datasets that differ by at most one individual), the output of the function $f$ satisfies: $\Pr\left[p\left(f(D)\right)\right] \leq\ e\ \varepsilon \Pr\left[p\left(f(D0)\right)\right]$ (1) for any potential outcome p. That is, the two probabilities cannot differ by a factor larger than $e\ \varepsilon$ [2]. Three core properties characterize DP:

(1) an output for an analysis must not compromise privacy;

(2) a crucial component of an analysis is a privacy-sensitive randomization process, where some noise is added to the output;

(3) the privacy guarantees do not depend on how well the adversary knows the dataset, or how many auxiliary sources of information the adversary has.

### 1.1. Definition and Principles of Differential Privacy

Differential Privacy (DP) is formally defined as a property of a subjective randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}|\mathcal{X}|is$ $(\epsilon, \delta)$ - deferentially private if for all S ⊆ Range($\mathcal{M}$) and for all $x, y\epsilon\mathbb{N}|\mathcal{X}|$ such that $||x - y||1 \leq 1$ will follow: (Formula), where the probability space is over the coin flips of the mechanism $\mathcal{M}$. Differential Privacy is inspired by the idea of maximizing the potential of public-interest data to help, while at the same time minimizing its corresponding risk to individual lives [1].

Core concepts in Differential Privacy (DP) are its privacy guarantees. According to its definition, even if an adversary knows nearly all other databases used by the same mechanism $\mathcal{M}$ and observes the output y, they should not be able to learn anything significant about the specific database, such as whether it is included or not. The term 'almost' refers to the fact that there are always some events with non-zero probability for which the given mechanism $\mathcal{M}$ ensures privacy, but the two databases may differ in their output probability distributions, potentially leading to significant risks [3].

## 2. The Need for Differential Privacy in Cybersecurity Analytics

Recent incidents involving data breaches targeting high-profile government agencies and businesses are indicative of a rising trend of advanced persistent threats (APTs). Threat-hunting teams play a crucial role in analytical activities involving the collection and examination of threat intelligence data concerning APT groups, malicious software (malware), and cyber-attacks. Traditionally, such data was shared as a wholesome file, but a burgeoning trend of sharing observational attributes by using Interactive Data Sharing Language (IDSL) language has emerged recently. Since this data contains sensitive intelligence, it is imperative to thoroughly examine its privacy-preserving methodology before employing it for collaborative intelligence analysis [1]

While academics have explored privacy-preserving data (PPD) techniques in academia and across numerous domains amply, there is no concurrent endeavor in the cybersecurity domain. By cybersecurity, reference is made to all activities surrounding the collection and analysis of threat data to build a shield against malicious cyber activity. Since such data sharing could suddenly leak sensitive intelligence, it fosters a motive to explore strong privacy-preserving techniques like differential privacy (DP). DP offers a strong mathematical promise of privacy, making it superior to many earlier PPD techniques [2]. Most existing DP data-sharing mechanisms are not congruous with IDSL-based analytics. Consequently, exploring such compatibility issues within analytics necessitates cross-laying prior knowledge from both analytics and DP perspectives, followed by suggestions to bridge the currently existing dichotomy therein.

### 2.1. Challenges in Preserving Privacy in Threat Intelligence Data

Organizations often seek to leverage threat intelligence from other sources to strengthen their cybersecurity defenses. However, sharing raw threat intelligence data can expose sensitive information, thereby increasing the risk of exploitation by malicious actors. Prior to sharing, organizations must redress data privacy concerns, which may include obscuring sensitive information such as IP addresses, user IDs, and other unique identifiers that link back to an organization’s internal environment [1].

Conventional measures for protecting privacy, such as data anonymization and obfuscation, may be inadequate. While anonymization can hide unique identifying attributes of the data, several studies have demonstrated that adversaries can still infer sensitive information through a combination of auxiliary knowledge and other identifying attributes (e.g., dates of birth, zip codes, etc.) [4]. Redacting specific values in the data may also be insufficient, as accurately guessing them may still lead to privacy violations. Through estimation approaches (e.g., using misspecified models), it is possible to make inferences about the redacted information. Additionally, shared datasets often have auxiliary datasets that can assist in inferring sensitive information (e.g., using a social network graph).

The notions of privacy have often focused on direct and specific privacy breaches, severely underestimating the information that can be leaked from the data. Therefore, there is a need for privacy-preserving mechanisms that provide rigorous privacy guarantees even when background knowledge is employed, and privacy is not compromised by the combination of several other pieces of information.

## 3. Applications of Differential Privacy in Cybersecurity

Security Information and Event Management (SIEM) systems are widely adopted to centrally analyze security data generated by a variety of devices and technologies. Typically, this involves deploying an agent on the device that collects and prepares the data before transmitting it to a SIEM infrastructure. In this context, differential privacy (DP) can be employed on the agents collecting events by adding noise to aggregation queries that analyze the events before forwarding the information to the SIEM back end [5]. This mitigates the risk of exposing sensitive individual information while still making it possible to derive significant insights from the data.

The promising results achieved in this research work indicate that it is feasible to integrate various forms of DP with existing SIEM tools, while satisfying constraints on the utility of the data. Reducing the risk of breaching personal information would also allow organizations to share their logs without fear of penalties. This is especially relevant in cases where it is desirable to collectively analyze statistics of different organizations' logs, e.g., to assess general threat intelligence. More generally, this would enable the distributed deployment of agents, which brings scalability gains.

### 3.1. Integrating DP with SIEM Systems for Secure Data Analysis

The surging frequency of cyber threats has led to the more rapid adoption of Security Information and Event Management (SIEM) systems, enabling enterprises to gather event data for analysis. An SIEM system is a unified process for managing system logs and security events that utilizes analysis and correlation to detect and alert the presence of aberrant activity [6]. The collected security data are stored in a database and provide the foundation for an analysis that is performed based on several views of the data, including user activity, access attempts, and file modification. However, the accessibility of sensitive security data in and outside an enterprise raises a serious concern regarding privacy and potential exposure. Cyber analysis is a powerful and extensible analytical method for big data that tracks queries and data quality and provides a mechanism for unconstrained analytics [1]. Big data fuels improvements in learning and comprehending data. Moreover, big data cannot be anonymized and released, as anonymity cannot provide strict privacy guarantees when the data is released. This has led to a new approach to big data protection based on differential privacy (DP), which provides privacy guarantees based on the inclusion or exclusion of any single record in the database.

Nevertheless, a mechanism for safely analyzing security data without jeopardizing privacy has not yet gained traction. Meanwhile, DP mechanisms often require extensive query constraints and can reduce the utility of the data being analyzed, which precludes complicated queries and exhaustive coverage in data analysis. Consequently, an SIEM system is integrated with DP for safe and efficient data analysis, where security data is fenced by a privacy-preserving data analyzer enforcing DP and ensuring that queries and updates to the data comply with the DP policy [7]. A fence query algorithm makes optimal use of DP-compliant OMIN operations, and a two-move approach for a reported mechanism ensures that sensitive security data are not exposed in any direct request to the SIEM system.

# 4. Balancing Data Utility and Privacy in Threat Intelligence

While there is no single, universally accepted solution to the problem of balancing privacy and utility in any particular application, social scientists and policy analysts need to better understand the complexity of decision-making in relation to trade-offs. There is a robust discussion of privacy-utility trade-offs in the literature on differential privacy, although such discussions are on the technical side, discussing the mathematical framework of trade-offs and privacy-loss budgets in terms of information theory [8]. Threat intelligence functionalists should empirically examine the balancing of privacy and security instrumental within the actual practices of cybersecurity analysts. From the perspective of cybersecurity analysts, data utility is understood as useability for cybersecurity analysis and intelligence production, flowing from consideration of both usefulness for a given analysis and veracity, quality, and precision. Threat intelligence is data or information that can be used to counter threats and vulnerabilities posed to the assets and information systems of organizations of all kinds. Within the trade-off, two areas of consideration arise – the privacy preserving techniques put into play and the impact upon data by implementing such techniques. From here, the most actionable goal is a trajectory towards optimal epsilon, or ε parameter. Indeed, this sensitivity region may be narrowed around a choice of ε for differentially-private mechanisms. *ε's* elasticity entails a privacy/utility trade-off [9]. If ε is large, the mechanisms have better utility from the accessibility of greater amounts of the data set. But privacy guarantees are weaker, as more of the unaltered data is fed to the algorithm—there is a non-linear relationship between ε and the probability of violation. Conversely, if ε is small, mechanisms provide greater privacy against external attacks, but lower data utility.

## 4.1. Optimizing Epsilon Parameter for Trade-offs

To balance the trade-off between data privacy and data usefulness, the epsilon parameter in differential privacy must be optimized. A mechanism's level of privacy is measured by a positive value called epsilon (ε); the lower the epsilon, the more stringent the privacy assurances, but the higher the data utility loss. On the other hand, a greater epsilon provides lower privacy but better data usefulness. [10]. The application's objectives and context must be taken into account in order to maximize Epsilon. This entails figuring out the epsilon value for cybersecurity analytics that protects people's privacy while optimizing the value of threat intelligence data. Adjusting epsilon entails testing with various values, assessing the effect on data utility (for example, using metrics like completeness or correctness), and making sure the privacy constraints as outlined by ethical and legal norms are fulfilled [8]. Performing synthetic tests with different epsilon values and examining the impact of noise generated by the Laplace or other processes on the data's analytical value are common steps in the optimization process. Organizations can select an epsilon that satisfies their privacy and utility requirements by weighing the trade-offs, finding a balance that preserves sensitive data while enabling efficient data use for security research.[9].

# 5. Case Studies and Practical Implementations

This section will detail real-world case studies and practical implementations that illustrate the use of differential privacy in cybersecurity analytics. By exploring tangible examples and experiences where differential privacy has been effectively employed, valuable insights will be gained into the concrete applications of differential privacy within the cybersecurity domain. This study will examine how differential privacy has been integrated into cybersecurity analytics practices, the challenges faced, and the impact of this integration. Through these case studies, the potential of differential privacy in protecting data while still enabling meaningful analysis will be demonstrated.

Cybersecurity analytics involves the examination of data across networks, servers, devices, and users to identify and mitigate risks. Organizations are increasingly adopting data-driven analytics as their primary strategy for dealing with cyber threats. However, in order to effectively analyze data for the purpose of cybersecurity investigations and to power detection algorithms, security analytics systems often rely on large volumes of sensitive and personal data. Data breaches of security analytics systems could have far-reaching implications for organizations, as well as harms for individuals. To mitigate privacy risks in such systems that analyze sensitive data, differential privacy has been explored as a solution [1]. Differential privacy allows data sharing with strong privacy guarantees, ensuring that the outputs of a query do not significantly differ with or without an individual's data in the data set.

### 5.1. Real-world Examples of DP in Cybersecurity Analytics

In cybersecurity, Differential Privacy (DP) is essential for doing data analysis while preserving individual privacy. By adding noise to data, DP helps Security Information and Event Management (SIEM) systems detect threats without disclosing private information [1]. By combining data in a manner that makes it impossible to identify particular entities, DP also makes it easier to share threat intelligence in a safe collaborative manner. Adding noise to statistical summaries in malware research helps create detection signatures and preserve individual data samples [11].

DP helps with user and network activity monitoring by protecting the privacy of behavior analysis while detecting any dangers. In incident response, DP assists with data analysis without jeopardizing the privacy of individuals [11]. Machine learning models may be developed on sensitive data while maintaining privacy thanks to the combination of DP with deep learning. Managing computing costs, maintaining regulatory compliance, and striking a balance between privacy and value are some of the difficulties. Notwithstanding these difficulties, DP's contribution to cybersecurity analytics is essential for preserving privacy and strengthening security protocols.

## 6. Evaluating the Effectiveness of Differential Privacy in Cybersecurity Analytics

The effectiveness of differential privacy (DP) in preserving privacy is evaluated using metrics and criteria from the perspective of threat intelligence analysis. Given raw data such as IP addresses used for cyber-attacks, reports on detected malware, and variants of collected malware, DP is applied to address general queries asked on the dataset, such as the number of queries in a specified time range. The relationship between the query and the privacy parameter is shown to reduce the sensitivity of the query counts, thereby preserving privacy. The output of these queries is appropriately perturbed to satisfy DP.

The effectiveness of DP is assessed under two aspects: (i) the likelihood of preserving privacy as intended after applying DP and (ii) the likelihood of success in gaining public knowledge from the output after applying DP. The average DP is used as the posterior privacy guarantee [12], and DP can be considered to preserve privacy. On the other hand, knowledge of the query and database provides a B-attack: the attacker estimates the effect of the database on the output of the query, and if the estimate is above a threshold, S will be deduced. A sufficient condition is formulated on the query count's privacy parameter such that the probability of success of the B-attack is below a certain threshold.

### 6.1. Metrics for Assessing Privacy Preservation

Different approaches have been implemented to assess the preservation of privacy in regard to differential privacy solutions. There are two main groups of metrics and parameters to assess the preservation of privacy in datasets. Metrics belong mainly to one of two categories: entropy based and re-identification based [13]. Both groups of metrics compute privacy scores from 0-1, but their interpretation is reversed since higher scores demonstrate better privacy in the first group and in the second group lower scores demonstrate better privacy.

Another function of the privacy related metrics is to compute the amount of protection that a privacy model or a particular anonymization method grant to the dataset. For a few popular privacy models, or classes of them, these metrics have been computed and they are published in articles. These metrics have been independently implemented and incorporated into the assessment framework to test the protection provided to the datasets. The efficacy of the metrics has been examined on a few benchmark datasets for different privacy preserving data publishing solutions [14].

## 7. Future Directions and Emerging Trends in Differential Privacy

While the previously discussed techniques of differential privacy in the Cyber Security domain address most of the critical emerging problems and challenges, there are many advancements and developments in the domain of such DP techniques that have a huge potential to help in shaping the future of this DP domain. Some of the most exciting and promising advancements/modifications/evolutions in DP techniques are discussed in this section which includes but is not limited to Adaptive Sample Size Selection Methods, Adaptive Mechanisms, Hybrid Mechanisms, Differential Privacy-Oriented Systems and Frameworks, Privacy-Preserving Hybrid Systems of Cyber Security, Applications of Deep Learning Framework in Protecting Cyber Security via Differential Privacy, and Emerging Threats and Defense Mechanisms for Differential Privacy in Cyber Security [1].

Different advancements/developments related to DP techniques can be implemented for malware data as Malware is one of the biggest threats to Cyber Security. Many advanced modifications in Cyber Security Malware threat analytics techniques have been proposed in the literature using two primary classes of techniques, namely, Evolving Computational Intelligence (CI) Techniques and Parallel Processing/Multi-Core Frameworks. In the future, DP could be used in combination with these existing advanced malware threat models to increase performance efficiency as well as ensure secure data disclosure. There are increasing quantifiable attacks on establishing differential privacy mechanisms. It is necessary to find out such types of attacks and the security design principles against those attacks. Researchers on differential privacy in the Cyber Security domain should focus on establishing robust differential privacy models which are unbreakable and secure to such types of attacks. There are also increasing attempts in Cyber Security to break differential privacy (DP) circumventing filters used to ensure DP. Research on robust Cyber Security filters against breaking differential privacy models may open new research directions in this area and can also ensure more secure comparative results [15].

### 7.1. Potential Advancements in DP Techniques for Cybersecurity

The cybersecurity sector recognizes the need to address privacy issues arising from the desire to share valuable data (such as threat intelligence), and uses various anonymization techniques (such as data suppression or noise addition transformation) [15]. This academic focus will probably widen DP to handle other domains and dataset types, such as non-Numerical data and relational datasets. So far, most DPSy works explore classic time, instance, and attribute level protections. Other dataset aspects offer exciting avenues for DP protection, such as complex graph type or unstructured High-Dimensional datasets. That will open a chance to explore the impacts of advanced DP protection types on cybersecurity analytics tasks. A noteworthy example would be the incorporation of Local-DP data protection [1] approaches within sensors.

Risk evaluation and privacy guarantees are two key aspects yet to be further explored on DP for all analyzed analytics tasks while uncertainty propagation and analysis in data analytic processes, as this task matrix shows many inconceivable effects. Another exciting opportunity lies in the study of the impacts of various attack strategies on DP-type guarantees to assess the privacy of the datasets. Tackling this opportunity can be done in conjunction with one of the former two avenues, which is to examine implications of conducting analyses on different strategies on the same data sets and compare results.

## 8. Conclusion and Key Takeaways

Threat intelligence (TI) is one of the main pillars of cybersecurity. It is a complex process that collects, analyze and shares data about threats and threat actors, to defend networks and assets from malicious activities [16] . However, this process comes with privacy concerns since some of the data being shared can be sensitive and critical to an organization. So far, most of the TI techniques proposed in the literature failed to properly protect the privacy of the data analyzed and shared. Recently, there has been an increasing interest in the adoption of differentially private algorithms on TI systems, which provide guarantees about what information is leaked. This paper analyzed the state-of-the-art in the integration of differential privacy mechanisms on TI systems by taking a closer look at the application of the proposed algorithms, their utility-privacy trade-offs and the protection they provide. Following that, it proposed a differential privacy mechanism for the analysis of billions of logs collected from an intrusion detection system. This mechanism was carefully designed to maximize the utility having in mind the privacy risks of deemed data. Additionally, it discussed the importance of selecting the proper privacy-utility trade-off for differential privacy mechanisms to be applicable. There is a challenge for organizations to adopt a privacy-preserving TI analysis with a reasonable sharing cost. In addition to this, the developed algorithms commonly do not specify the privacy-utility rate, making it impossible to know if it is well tuned and applicable on real scenarios [1].

A proper privacy-preserving TI analysis can unlock the potential of sharing logs from different organizations, increasing the visibility of emerging threats. This would make it harder for attackers to perform reconnaissance and have a better understanding of the defended environment. It would offer realistic defense solutions for low-budget companies and prevent the widespread of malicious tools. On the other hand, it is important to mention that adopting differential privacy mechanisms does not eliminate the need for already proposed TI privacy-aware systems. There are many ways that private information can be leaked, and differential privacy guarantees do not provide protection against every possibility [17]. Thus, it is a collective approach that should be adopted in conjunction with other privacy protection mechanisms.

## References

[1] P. Sengupta, S. Paul, and S. Mishra, “Learning with differential privacy,” in *Handbook of Research on Cyber Crime and Information Privacy*, IGI Global, 2021, pp. 372–395. Accessed: Dec. 01, 2024. [Online]. Available: https://www.igi-global.com/chapter/learning-with-differential-privacy/261739

[2] D. Leoni, “Non-interactive differential privacy: a survey,” in *Proceedings of the First International Workshop on Open Data*, Nantes France: ACM, May 2012, pp. 40–52. doi: 10.1145/2422604.2422611.

[3] H. Ebadi, *Dynamic Enforcement of Differential Privacy*. Chalmers Tekniska Hogskola (Sweden), 2018. Accessed: Dec. 01, 2024. [Online]. Available: https://search.proquest.com/openview/3a02ed3ff074cddbd4c9fab43fa0e5a1/1?pq-origsite=gscholar&cbl=18750&diss=y

[4] S. Sousa, C. Guetl, and R. Kern, “Privacy in Open Search: A Review of Challenges and Solutions,” Apr. 04, 2022, *arXiv*: arXiv:2110.10720. doi: 10.48550/arXiv.2110.10720.

[5] G. González-Granadillo, S. González-Zarzosa, and R. Diaz, “Security information and event management (SIEM): analysis, trends, and usage in critical infrastructures,” *Sensors*, vol. 21, no. 14, p. 4759, 2021.

[6] S. M. Zeinali, “Analysis of security information and event management (SIEM) evasion and detection methods,” *Tallinn University of Technology*, 2016, Accessed: Dec. 01, 2024.

[Online]. Available: http://mendillo.info/seguridad/tesis/Morteza.pdf

[7] J. Wei, E. Bao, X. Xiao, and Y. Yang, "DPIS: An Enhanced Mechanism for Differentially Private SGD with Importance Sampling," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles CA USA: ACM, Nov. 2022, pp. 2885–2899. doi: 10.1145/3548606.3560562.

[8] P. Nanayakkara, J. Bater, X. He, J. Hullman, and J. Rogers, "Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases," Jan. 16, 2022, *arXiv*: arXiv:2201.05964. doi: 10.48550/arXiv.2201.05964.

[9] Y. Li, Y. Liu, B. Li, W. Wang, and N. Liu, "Towards practical differential privacy in data analysis: Understanding the effect of epsilon on utility in private erm," *Computers & Security*, vol. 128, p. 103147, 2023.

[10] S. H. Alkaabi, "VISUALIZING PRIVATELY PROTECTED DATA: EXPLORING THE PRIVACY-UTILITY TRADE-OFFS," 2024, Accessed: Dec. 01, 2024. [Online]. Available: https://scholarworks.uaeu.ac.ae/all_theses/1203/

[11] R. Cummings, S. Hod, J. Sarathy, and M. Swanberg, "ATTAXONOMY: Unpacking Differential Privacy Guarantees Against Practical Adversaries," May 02, 2024, *arXiv*: arXiv:2405.01716. doi: 10.48550/arXiv.2405.01716.

[12] S. Zhang, A. Hagermalm, S. Slavnic, E. M. Schiller, and M. Almgren, "Evaluation of Open-Source Tools for Differential Privacy," *Sensors*, vol. 23, no. 14, p. 6509, 2023.

[13] J. Domingo-Ferrer, K. Muralidhar, and M. Bras-Amorós, "General confidentiality and utility metrics for privacy-preserving data publishing based on the permutation model," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 5, pp. 2506–2517, 2020.

[14] Y. Zhao and I. Wagner, "Using metrics suites to improve the measurement of privacy in graphs," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 1, pp. 259–274, 2020.

[15] D. Heredia-Ductram, M. Nunez-del-Prado, and H. Alatrista-Salas, "Toward a comparison of classical and new privacy mechanism," *Entropy*, vol. 23, no. 4, p. 467, 2021.

[16] B. Nour, M. Pourzandi, and M. Debbabi, "A survey on threat hunting in enterprise networks," *IEEE Communications Surveys & Tutorials*, 2023, Accessed: Dec. 01, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10216378/

[17] P. Kairouz, "The fundamental limits of statistical data privacy," PhD Thesis, University of Illinois at Urbana-Champaign, 2016. Accessed: Dec. 01, 2024. [Online]. Available: https://www.ideals.illinois.edu/items/95105

# Hybrid Multi-Factor Authentication (MFA) Using Biometrics and Behavioral Analysis

BOUFAIDA SOUNDES OUMAIMA[1], BENMACHICHE ABDEMADJID[2] and MAATALLAH MAJDA[3]

[1] *Faculty of Technology,Departement of Computer Science Universty of Chadli Bendjedid El Tarf, Algeria*

*s.boufaida@univ-eltarf.dz*

[2] *Faculty of Technology,Departement of Computer Science Universty of Chadli Bendjedid El Tarf, Algeria*

*benmachiche-abdelmadjid@univ-eltarf.dz*

[3] *Faculty of Technology,Departement of Computer Science Universty of Chadli Bendjedid El Tarf, Algeria*

*maatallah-majda@univ-eltarf.dz*

## Abstract

In the modern digital landscape, protecting identities and securing access to sensitive information has become increasingly critical due to the growing risks of data tampering, theft, and impersonation Multi-Factor Authentication (MFA) is an effective solution when numerous authentication methods are integrated This study introduces a Hybrid MFA system that integrates facial biometrics with behavioral analysis to enhance web service security Taking use of the unique and reliable characteristics of physiological biometrics, the system uses facial recognition as its first line of defense using Convolutional Neural Networks (CNNs). To counteract potential vulnerabilities such as spoofing, a second layer evaluates Keystroke Dynamics (KD), which captures unique typing behaviors, including dwell time and typing speed. The shortcomings of single-factor authentication methods are addressed by this two-layered strategy, which strikes a balance between user ease and security. The sequential verification process begins with facial recognition, followed by behavioral analysis, creating a reliable framework for controlling access. By merging advanced deep learning techniques with behavioral biometrics, this system offers an innovative solution for minimizing password-related security breaches and improving identity verification in online environments. The findings highlight the value of combining physiological and behavioral biometrics to strengthen authentication processes, reduce unauthorized access risks, and tackle challenges associated with privacy, spoofing, and data security.

**Keywords:** Multi-Factor Authentication (MFA), Facial Biometrics, Behavioral Biometrics, Keystroke Dynamics (KD), Convolutional Neural Networks (CNNs), Identity Verification, Cybersecurity, Deep Learning, User Authentication, Access Control Systems.

## 1. Introduction

Today, many people perform daily tasks requiring identification on computers or the Internet. Giving computerized identity information and controlling access to them is critical in this digital world. But computerized subject and identification information security is vulnerable to tampering, stealing, and imitation risks. Multi-Factor Authentication (MFA) is a technique for establishing and confirming an identity using two or more diverse authentications. MFA can be classified as a hybrid of two individual authentication categories. This paper proposes a Hybrid Multi-Factor Authentication (MFA) using a combination of Biometrics and Behavioral Analysis to facilitate secure authentication through a comprehensive approach involving two or more authentication techniques working together, thus improving security for unauthorized access to sensitive or personal information [1]. Due to their uniqueness (to a person), a biometric trait, and a behavioral biometric technology, MFA is proposed. The major advantage of biometric MFA is that, unlike traditional password, token, and card-based security systems, biometric traits are not easy to steal and replicate. Hence, the chances of hacking and unauthorized impersonation of individuals are minimal. However, no biometric trait is perfect. In real-world applications, there are always risks of false matches, false non-matches, and even a perfectly matching biometric trait can lose its uniqueness over time [2]. A combination of behavioral biometrics and traditional biometric technology is proposed using a hybrid approach to overcome all these risks and vulnerabilities and facilitate accurate and reliable authentication.

### 1.1. Background and Significance:

In most computer systems, user authentication is considered a fundamental building block and a primary line of defense for computer security [3]. Furthermore, with the requirements, versatility, and robustness of the authentication process constantly growing, it is important that this process offers a balance between the security of the system and the usability for the users. Nevertheless, the lack of flexibility and user-centricity in the current authentication methods subjects the users to a risk of either disclosing personal data or facing constant security breaches. The same outcome can be expected if the proposed future solutions heading towards a total abandonment of passwords – be it a sole biometric based or a hardware-security focused method – are fully implemented in the current systems, as they rely too much on either one factor of authentication or a strongly fused method anchored on an external device. Special care ought to be taken to limit the scenarios wherein an attacker would have the same possibility of a user (according to the time, location, and user's features) to access the system, and thus make it unreasonably challenging to achieve.

Physical biometric information (such as fingerprinting, handprints, iris and retinal scanners, and face recognition) as well as behavioral and cognitive biometric data (such as voice, keystroke dynamics, surface dynamics, gait, and signature) of a user are sometimes back-end stored by enterprises providing systems, devices, applications, or services to its user(s). Implicitly, the user grants permission and agreement that their biometric data becomes property of the enterprise in order to access such systems and services [1]. Concerns arise later however, regarding the limited control, confidentiality, data portability, and exploitation of their own biometric data, as well the consequences on privacy and security of breaches that could lead to unwanted outcomes. On a different point of view, within the current conceptual framework prioritizing security and privacy, the user may also challenge the trustworthiness of the systems, devices, applications, and services alike, as many have already been victims of sophisticated deception, forgery, and manipulation attempts.

### 1.2. Research Objectives:

To reduce password-related intrusions, a Hybrid Multi-Factor Authentication (MFA) system for web services using facial biometrics and behavioral analysis is proposed. Biometric authentication techniques can automatically identify or verify a person by evaluating one or more distinguishing biological traits and can be classified into physiological biometrics and behavioral biometrics [1]. Given the advancements in deep learning, facial traits seem to have gained significant prominence in biometric systems. However, given the possible spoofing attacks, an effort has also been made to consider behavioral biometrics simultaneously.

In the behavioral biometrics domain, keystroke dynamics are generally the most widely used web- and text-based authentication technique. It is based on the unique typing behavior of individuals using factors such as dwell time, flight time, typing speed, etc., captured through key logs on keyboards, virtual keyboards, or touch devices [3]. A multi-stage two-factor security system based on the combination of facial biometrics and keystroke dynamics is proposed here. The first line of defense is the face authentication using the deep learning-based Convolutional Neural Network (CNN) classification model with the input of 2D facial color images. The second line of defense is to check the identity-given typing behavior of individuals using Keystroke Dynamics (KD). A sequential system architecture is built where at first face-authentication is performed and on its positive acceptance a KD-based authentication model is considered.

## 2. Fundamentals of Multi-Factor Authentication

Multi-Factor Authentication (MFA) is a widely deployed security measure to curb cyber-attacks across web applications. MFA provides enhanced cyber safety by combining the benefits of multiple authentication techniques which a user must satisfy in order to gain access or authenticate to a target system. With a single layer of security, if that one credential is compromised, the account is at risk. As internet security measures became more complex, the cyber criminals have also become more tech savvy to exploit loopholes in security mechanisms. In addition to passwords and credentials, multi-factor authentication requires conditionally independent ribbons of authentication (identifiers) for improving cyber-security and privacy [4]. There are several types of such identifiers which can be used for achieving authentication: (1) Tokens, Cards and Devices – this includes all kinds of hardware which issue security codes; (2) Knowledge-based passwords – this includes PINs or responses to secret questions; (3) Biometric authentication – this is based on physical characteristics of the user and includes fingerprint, iris and voice-recognition.

Users tend to favour web applications which utilize simpler methods for Security Reasons. Therefore, it can be difficult to introduce State-of-the-Art security measures on pre-existing applications, as this can frustrate a user's experience and many users may choose to stop using the application [5]. On the other hand, in caso of possession-based systems, most devices may not support authentication systems which are resource hungry or very complex for the device in question. Thus, although State-of-the-Art multi-factor authentication systems can be extremely potent in terms of accuracy of verification and robustness, the implementation for existing applications can be very hard or even impossible. It is noticed that certain classes of flaws appear across classes of web applications employing similar measures for security. Cyber safety is improved by diversifying the devices used for authentication as to satisfy a broader set of conditions of independence.

### 2.1. Definition and Importance

Multi-Factor Authentication (MFA) is a security measure in which the user is granted access only after satisfactorily presenting two or more pieces of evidence to an authentication mechanism. The second factors in a multi-factor authentication system can fall primarily into one of three categories: "something you know," such as a password or personal identification number (PIN); "something you have," such as a passcode-generating device, smart card, or security token; or "something you are," such as a fingerprint, iris or face recognition, or voiceprint. Risk-based authentication falls outside these factors, but it has also gained some traction in the user authentication literature [4]. Multi-Factor Authentication is particularly important in protecting sensitive or valuable information, including confidential medical records or corporate data. Passwords present a poor defense in depth, as they are often the only mechanism protecting access to sensitive information. Passwords can be stolen through brute force attacks, phishing, keylogging, password cracking and social engineering. If stolen, a password can be used not only to steal sensitive information but also to impersonate the victim. A single password can therefore represent a single point of failure in a security system.

Multi-Factor Authentication addresses the weaknesses associated with passwords. Even if an attacker steals a password through a phishing attack, it would be useless unless the attacker also has possession of the second authentication factor, making the theft of a password alone not sufficient to impersonate the victim. Implementing MFA methods is crucial in preventing unauthorized access to sensitive information. This is clearly illustrated by the IBM Security Services 2015 Cyber Security Intelligence Index, which states that “90 percent of security incidents involve stolen or weak passwords” [5]. A multi-factor solution has therefore become an essential part of any security system that protects sensitive information. Because sensitive information can be stolen, impersonated or passed on from one person to another, the importance of MFA methods becomes obvious.

## 2.2. Traditional MFA Methods

These are the simplest and most traditional forms of authentication. Users provide a unique username and a password (usually a secret word or string of characters) to access a particular system. However, security can be compromised if the password is stolen or guessed [4].

A security token is a physical object that the user possesses to authenticate their identity. It includes smart cards and USB tokens that generate one-time passwords (OTP). Security tokens enhance security but may be misplaced, thereby hindering access.

This uses the unique physical characteristics of users for verification and access control. Common biometric traits include fingerprints, voice, face, iris, and keystroke. While biometric systems are unique, unchangeable, and easy to use, they can be spoofed and are more costly to implement.

# 3. Biometric Authentication

The biometric modality involves a person’s physical or behavioral characteristics for identity verification. Large data storage management, security and privacy risk are the key challenges within biometric authentication systems. Biometric traits can be categorized into two main categories: physiological and behavioral traits [1]. Physical biometric modalities including face, fingerprint, hand geometry, hand vein, palm print, iris, ear shape, voice-sound, DNA, gait, and retinal. Behavioral biometric modalities related to mental status including keystroke dynamics, signature verification, online hand writing, smile and rhythm of speech.

With the recent advancements in artificial intelligence, deep learning technology has extensively been adopted in biometric recognition applications. Facial recognition by CNN-based deep learning training and thumb notch recognition by the irregularity detection method is implemented. CNN-SVM architecture using deep learning for comparison of recognition performance of gender biometric tool with soft computing gender detection is proposed. Different gait features like joint angles, foot-off, foot-contact, position and phase sequences are extracted to identify the dynamic signature of each gait motion. The surface electromagnetic field, characteristics of the current, the potential difference and the signal spectrum are used to decompose the signals into high dimensional feature vectors.

## 3.1. Types of Biometric Modalities

From the variety of biometrics, the following types and categories of biometric modalities are used for authentication [1]. Each biometric modality is briefly described:

Facial Recognition Systems: Facial recognition is the ability of a system to detect and recognize human faces. It utilizes the unique characteristics of facial images to distinguish one person from another. It is one of the most commonly used biometrics for authentication because it can be executed without the individual's knowledge. For example, airports utilize facial recognition to compare surveillance video and photographs of faces in the database. A facial recognition system generally consists of four modules: face detection and tracking, facial feature finding, face representation, and matching. Fingerprint Recognition Systems: among physiological biometrics, fingerprinting is the most widely used biometric method. The knowledge of fingerprinting is as old as the five-fingered hominoids. Fingerprints are unique characteristics of every human being. Fingerprints considered biometrics are ridge patterns present in the skin. Based on ridge patterns, there are three classes to cluster fingerprints: loop, whorl and arch. Recognition can be achieved either through matching the whole fingerprint image or matching selected minutiae points attached with fingerprint features. Hand Geometry Recognition Systems: Hand geometry is a very well-established biometric. It is a mature biometric technology that can recognize individuals by their hand shapes. This biometric can also be used with fixed terminals at the entrance gate to check authorization. Individual authentication based on hand geometry is performed using various features such as palm area, folded surface features, palm texture, etc. The palm texture is used to recognize individuals using either statistical or learning techniques. Iris Recognition Systems: In an iris recognition system, a circular piece of an image with the eyeball (iris) boundary is captured using a camera. Iris corresponds to the pupil of an eye in which color and texture patterns are present. This texture pattern is unique, consistent, and stable, making it a mature biometric for individual recognition. Various algorithms have been proposed for iris recognition, and results have been published concerning their efficacy. Voice Recognition Systems: Voice recognition is a behavioral biometric that attempts to authenticate individuals based on their voice patterns. Voice quality varies on three different levels: physiological, psychological and sociolinguistic. The speaker's voice is acquired from a microphone, and it is digitized and analyzed in voice recognition. The digitized voice goes through different processing steps to extract voice features and a matching score is computed. Signature Recognition Systems: A handwritten signature is considered a behavioral biometric that attempts to model pattern identification technique to recognize individuals based on signature dynamics such as trajectory, velocity, and acceleration. Signature authentication could benefit any application in which the user signs a transaction, e.g., financial or legal contracts. Signature dynamics reflect personal habits in how an individual signs his/her name, relating to pen pressure, speed, and stroke order. Vein Recognition Systems: Vein recognition is a physiological biometric that attempts to authenticate individuals based on vein patterns such as palm vein and finger vein. Vein biometrics are used for security purposes since vein patterns are embedded in human soft tissue. Fluctuations in vain patterns caused by aging or injury are negligible, making it a mature biometric for individual authentication.

## 3.2. Advantages and Challenges

A biometric authentication system identifies an individual based on their unique physiological or behavioral characteristics. These characteristics could include fingerprint scans, voice recognition, facial recognition, retina scans, keystroke dynamics, etc. Over the past decade, many nations and organizations have integrated biometric authentication systems into sensitive applications. In these applications, traditional authentication methods could not hold against everyday threats of impersonation, collusion, hacking, theft, or plague. Hence, the resurgence of biometrics [1].

Biometric characteristics are based on anatomical or behavioral features that are permanent, consistent throughout life, difficult to forge, and specific to an individual. The anatomy or behavioral features which are the basis of biometric systems could be either physiological features such as face, fingerprint, iris, voice, etc., or behavioral features such as gait, signature, keystroke dynamics, etc. The most significant advantage of biometric authentication over traditional systems where authentication is based on knowledge (e.g., passwords or pin numbers) or possessions (e.g., tokens or smart cards) is that biometric characteristics cannot be copied, borrowed, or forgotten.

The advantages of biometric authentication are usability: user cannot forget their face, iris or fingerprint, and Once biometrics are enrolled, there is no need to change them unless some catastrophic damage is done. Specificity: Except for the recently proposed theories, there are no two known individuals with the same biometric characteristics. The disadvantages of biometric authentication are attacks: forged biometrics, convergence of two biometric signatures, intra-class variations, tracking/following of individuals, impersonation, deep fakes. Privacy: unintended access to sensitive biometrics and concerns relating to people's nature and habits (thoughts, intention). Social discrimination/prejudice: race, gender, incapacity. Protection of minors: restricting the access to sensitive spheres [6].

# 4. Behavioral Analysis in Authentication

Human behavior is considered a distinctive characteristic of an individual that can be beneficial for addressing various security concerns. Behavioral biometrics, which involves recognition based on behavioral features (e.g., mouse and keystroke dynamics, touch, gait analysis, or voice), is for identity verification with a particular emphasis on continuous authentication. A proper deployment of biometrics can substantially safeguard resources against unauthorized access [5]. Behavior-based biometrics is less intrusive than other biometrics, such as facial or speech. Although the physical structure is constant, a user's behavior is dynamic and can vary based on factors such as device usage, workload, and environmental conditions. Continuous behavioral biometrics offers a way to diminish security concerns without imposing overt passwords and strict concurrent user authority. Traditional authentication refers to a one-shot decision based on the initial login, whereas behavioral biometrics continuously updates an ongoing model of the user and makes timely decisions based on that model. Behavioral analysis generally includes complementary biometric traits (such as mouse dynamics). The goal is to assist in the performance and robustness of the biometric system by utilizing available data while achieving the desired trade-off between accuracy and computational cost. This method can also filter out uninformative or irrelevant features by measuring their individual contribution to the task [7].

With the rapid evolution of communication technologies, users can access various online services or sensitive information using a multitude of computing devices (e.g., laptops, tablets, and smartphones). Passwords and similar conventional means are insufficient to protect identity and access authorization. Combining two or more credentials is known as multi-factor authentication (MFA). There are two principal categories of MFA: knowledge-based and biometrics. A combination of the two can yield a hybrid MFA system. Knowledge-based MFA comprises multiple credentials that a user knows and can sufficiently provide in the authentication process. A classic example is a password or PIN along with a phone call or security question. The weakness of this methodology lies in the memorization of credentials and possible sharing, covert use, or the tendency to use easily guessed PINs. A biometric approach cannot be impersonated or forgotten; it includes physical or behavioral characteristics of a user. The former consists of morphological traits (fingerprint, face, hand geometry, iris, or retinal scans), while the latter includes voice, keystroke, medical, or gait features. The strength of biometrics comes from the difficulty to forge or share and the uniqueness of traits among users.

## 4.1. Types of Behavioral Biometrics

Gait recognition is particularly useful for continuous authentication. As a user walks, the mobile device being carried may experience repetitive, periodic, and consistent behaviors resulting from the user's gait and body movement. Handwriting dynamics capture strokes written with a stylus, which can be used for user verification and also indicates user attention. Signature dynamics capture strokes on touch devices, requiring only low-cost hardware. Wireless signals can capture a user's physiological and behavioral features. Physical features include height, age, speed, and weight, while behavioral features include walking speed, gait stability, and movement patterns in different locations. Lip movement capturing vocal and non-vocal features can model an individual's unique communication behavior [7].

Behavioral biometric systems have been explored extensively in addition to physical biometrics. Individuals have a reasonable expectation of continuous or periodical authentication throughout their daily use of an IT system. Behavioral biometrics, notably keystroke dynamics and mouse movement, can provide forms of implicit authentication, neither of which requires any effort from the user. Behavioral biometrics focused on a single biometric or exploit context information to improve authentication decisions have been researched [8].

## 4.2. Benefits and Limitations

The concept of using behavioral biometrics for authentication is not entirely novel. Research relevant to behavioral biometrics and its security implications has been extensively documented [7]. Behavioral biometric systems have been adopted to augment authentication mechanisms by leveraging human behavioral characteristics, such as gait, keystroke dynamics, and mouse movement patterns. Physical biometrics typically requires dedicated and/or expensive equipment such as cameras and/or fingerprint scanners, making them more difficult, if not impossible, to deploy in some scenarios. In contrast, behavioral biometrics utilize existing deployed hardware on devices, making implementation easier. Among all the behavioral biometrics, behavioral biometrics based on mouse movement arguably has the most common form of implicit authentication simply due to the ubiquitous use of the mouse. Mouse movement patterns are intrinsic to each individual, and it has been shown that mouse movements reveal consistent patterns among individual users.

However, there are inherent drawbacks of leveraging behavioral biometrics for authentication. As alluded to in earlier discussions, behavioral biometrics is susceptible to external environmental changes. Substantial changes in environment often cause performance degradation, especially in continuous authentication scenarios. Extensive additional effort is often required to maintain stable performance levels to account for changes in environment. In addition, there is considerable participant variation with respect to the characteristics of behavioral biometrics. Such variability can hinder the generalization capability of a behavioral biometric system across different users, as a behavioral biometric model for one user is often ineffective for another user. Numerous attempts have been made to address both limitations, but little is known about the appropriateness of using behavioral biometrics in specific situations [5].

You should use the pre-defined styles for sections (Heading 1), subsections (Heading 2), and subsubsections (Heading 3).

There should be no empty lines before section headings. The template already adds the necessary spacing before them.

# 5. Hybrid MFA Concept

Multi-Factor Authentication (MFA) is based on the premise that an individual has to present two or more distinct types of authentication factors in order to gain access to resources or services. Authentication factors have traditionally included knowledge-based, possession-based, or inherence-based factors. There is wide acceptance of a Multi-Factor Authentication framework composed of two or more factors, as this enhances the robustness of the authentication framework itself. Various techniques are employed to authenticate users during different authentication phases. Such familiar authentication techniques include passwords, PINs, patterns, tokens, cards, fingerprints, irises, and voices. The focus is on solutions that combine different factors to authenticate individuals. A novel approach is suggested, which is an aggregate of biometric and behavioral analysis for authentication using smart devices. The system incorporates four factors for authentication, of which two are biometric-based factors and two are behavioral-based factors. Biometric-based factors use characteristics that address inherence types, employing the voice and keystroke pattern. Whereas behavioral-based factors are knowledge and possession types, the former requiring the user to answer small questions and the latter confirming the possession of one device [5].

Biometric authentication systems have attracted substantial attention due to their convenience and high security levels. Biometrics inherently addresses the "what you are" authentication approach, a type of factor that is intrinsically linked to an individual and is regarded as difficult to duplicate or share. There are some biometric traits that have gained in popularity and are widely adopted due to their extensiveness, consistency, and robustness. This includes fingerprints, iris scanning, facial traits, and voice recognition pattern. Behavioral-based biometrics employ implicit manners and habits of individuals, distinguishing people using their ordinary behavior. Every user interacts with applications in a unique manner that demonstrates a habitual pattern particular to him/her. Behavioral-based biometrics can be both explicit and implicit. The first deals with conscious actions, like signature verification, where people have to produce reviews at every access. Implicit-based behavioral factors authenticate users with no-interaction, or no-effort-required techniques, in which the detection and recognition happen unnoticeable to their activities [4].

## 5.1. Definition and Components

A system is developed to conduct Hybrid Multi-Factor Authentication (MFA) by combining Biometrics and Behavioral analysis. Hybrid-authentication ensures authorized access by correlating the identity of the user with her/his physical traits and behavioral traits. Physical traits are identifiers that are innate to an individual and cannot be dynamically altered. These traits are graphical in nature and include fingerprints, retina, face etc. On the other hand, behavioral traits are based on the user's actions over a period of time and include keystroke dynamics, touch dynamics, user movement patterns etc. User's movement patterns can be understood by observing a user's accelerometer and gyroscope data. Analysis of this data using sensor mining techniques can generate behavioral traits including acceleration, pose, speed and angle of rotation which constitute the motion-based biometric model [5]. The Hybrid Multi-Factor Authentication (MFA) model presented in the study comprises two components – (1) Biometric model (2) Sensor-based Behavioral model. The biometric model authenticates a user based on the facial features extracted from the user's face image while the behavioral model authenticates a user based on the movement patterns of his/her smartphone during a call. The behavioral features concerning the Phone Speed, Swings Per Call, deviate from the Initial Point etc. are extracted from the sensor data during calls.

The supporting architectural framework of the Hybrid Multi-Factor Authentication (MFA) system is depicted in Figure 1. The model is designed to work in a Cloud-based environment where all users' data is stored in a central database on the Cloud and the authentication processes are executed using the data stored in the Cloud. Any user wishing to join the Hybrid-MFA system needs to register for the service by providing her/his basic details along with their legitimate Face image & Mobile IEMI number for identification followed by the logging of her/his Call data for a specified period of time (minimum 1 week). Once the Call data is received, the sensor features are extracted and modeled at the server-side. The User's data during registration is stored in the Cloud in the User DB table using the registration service. Then by invoking the 'Auth_MFAservice', the user can authenticate to access the secured resources using Hybrid-MFA.

## 5.2. Advantages Over Single-Factor Systems

While single-factor authentication systems, such as passwords or tokens, are commonly employed by many organizations as a first line of defense, they are relatively easy to compromise due to social engineering, phishing, and hardware vulnerabilities. As a result, sensitive data is often stored in unencrypted formats even after being deemed high-risk. Presently, there are a number of solutions that provide two-factor authentication using text messages (SMS) and one-time passwords (OTP) sent to users' (token holder) mobile phones. However, several vulnerabilities are associated with these solutions, such as the chance of phone number compromise. A hybrid model is proposed in this study to enhance both authentication accuracy and security aspects. This model consists of verification based on audio and video signals, which can be performed using inexpensive hardware such as web cameras and microphones. The audio and visual biometric data is very dependent on a user's behavioral tendencies and is, therefore, unique to each individual. During the verification process, the audio and visual features of the data models of the authorized user are compared in real-time with the features of the inquiry data [5]. The presented model is a unique combination of various biometric systems that complement each other, and due to the nature of the systems involved, it is difficult to forge them without arousing suspicion. Additionally, it is common practice to change passwords periodically due to security concerns, which necessitates a new authentication model every time the password is changed. This is not the case for the proposed hybrid model. The strength of network security is determined by the weakest link, and any backdoor to the system may compromise the entire security measure [4]. The proposed hybrid biometrics system is more secure than a token-based system for two reasons. First, it does not allow unauthorized persons to appear as token holders, and second, it uses different biometric characteristics that could not come from a single person, making it impossible to forge these characteristics without arousing suspicion.

## 6. Integration of Biometrics and Behavioral Analysis

The integration of biometrics and behavioral analysis is a sophisticated endeavor that necessitates navigating several hurdles. First, the two systems typically employ disparate programming languages and architectures. In this instance, the behavioral analysis is designed using JavaScript, while the biometrics employ a Python-based Django framework. Second, both need to be operated independently but connected simultaneously. This ensures that when one verification is carried out, it does not halt the other from continuing. This presents an interesting challenge regarding how to synchronize two systems seamlessly with minimal stop time for the user. Third, the behavioral analysis verification occurs in the initial stage, while the biometric verification occurs later. As a result, there is a need for a robust queuing system that can handle an influx of users effectively without any build-up or delay for individual users [6].

The first challenge can be surmounted by utilizing a combination of MySQL and Redux. A common database shared by both systems can be built under the MySQL Connector module. This can then be called on both the Django and JavaScript sides, allowing either to pull or push data from the same pool. Alongside this, Redux can be used to manage the front-end state of the behavioral analysis application. Connected to the database, this can sense when the data is accurately inputted onto the MySQL side and display a pop-up to the user on the JavaScript side [9]. This means that timing is adjusted on both sides to the same value, guaranteeing that the two verifications are sequenced correctly. Regarding the second challenge, asynchronous programming is employed on the Django side. Through this, the biometric verification can be launched aside from the main workflow, permitting the behavioral analysis input to be processed and output prior to any halting at the end.

### 6.1. Challenges and Solutions

Although multi-factor authentication systems embedding both biometric authentication and behavioral analysis authentication exhibit great potentiality, some challenges addressing computer science and security issues may arise due to their complexity. Most notably, five major challenges can be identified.

As multifarious biometrics and behavioral features can be utilized in this hybrid MFA system [3]. Each system must have key insights regarding its usage conditions (e.g., for university environments, touch devices, smart devices, or PCs? For gait or motion features of smart devices in a vehicle, public transport, or walking at low speed? For users with disabilities?). These limitations or conditions will affect the implementation and performance of the hybrid MFA. Effective MFA system design can be based upon the underlying knowledge of the limitations, conditions, and pros/cons of both biometric and behavioral sources [3]. Thus, the key insights directing the design of such a hybrid MFA need to be analyzed.

The design of a practical hybrid MFA system must aim to tackle the question of how the combination of biometric authentication and behavioral analysis authentication approaches can complement one another. Beyond the idea of just synthesizing scores for decision-making, both authentication approaches must be well addressed to present their strengths and weaknesses. This aim is especially relevant since recent research suggests that fusing heterogeneous biometric and behavioral sources always achieves better performance than composing homogeneous sources. Different types of biometric and behavioral sources typically extract information associated with different patterns for discrimination [5].

For the algorithm's development, classification attacks against both biometric modalities (with synthetic traits spoofing the physiological biometric modality) and behavioral modalities (normal/abnormal behavioral traits) must be well consistent to explore the vulnerabilities of the hybrid MFA. Since some of these algorithms can be developed with publicly available datasets, this vulnerability exploitation is an important complement to the proposed development of a secure hybrid MFA with the initial design.

### 6.2. Algorithm Development

The primary algorithms are described in this section that allows the hybrid MFA to smoothly integrate biometrics and behavioral analysis. The overall workings of the hybrid MFA is broken down into two parts: first, the data acquisition and hybrid MFA algorithm operation on biometrics and behavioral analysis for a single user as one stand-alone system; and, second, the integration of the two into the overall hybrid MFA operations. Technical considerations of the algorithms involved within the hybrid MFA are explained. The pseudo-code for the algorithm development is provided as follows.

Biometrics – Touch Sensor Arc Length Algorithm: This subsection describes a continually running Biometric User Authentication algorithm based on the arc lengths of touch movements on the screen. The algorithm is briefly explained with the following steps: 1. Take 400 touch movement records on the screen and compute the movement arc length for each of the records. 2. Mean center the computed arc lengths to a common average arc length per record transformed to 400 Arc Weighted Lengths (AWL) numbers. 3. Use a weighted sum of First Order to Fifth Order AWLs to compute User AWL Bio and then store it for future comparisons. 4. When a touch movement is detected on the screen, compute the Movement AWL Record (1-5) and an AWL User Vector of that movement. 5. Compare the AWL User Vector (1-5) against the User AWL Bio mean center on 5 Arc Length Ratio Difference (ALRD) values. 6. Compute an overall score/student ratio of 5 ALRD values (DiffMax) and set pass/fail criteria less than defined threshold or greater. 7. Let each character's output (COM output '1' for authentic or '0' for inauthentic) be sent to a voting unit (COM output 5 steps total if '1' and 0 steps if '0'). 8. If majority voting output is greater than two, then authenticate user password character; otherwise, reject user password character [10].

# 7. Applications of Hybrid MFA

Banking and Finance: The banking and finance sector have been a pioneer in using MFA and have implemented security protocols and policies across all possible channels [4]. Traditionally, banks have relied on a secure user id-password combination to prevent unauthorized access to their accounts, but many banks now use one-time passwords (OTPs) through SMS or email alongside this. High-value transactions may demand further steps, such as biometric authentication via a fingerprint reader or iris scanner. However, this high-level security solution is not currently utilized for every transaction due to additional investment and customer convenience. As a part of its Operation Cyber Safe initiative, the Security Exchange Board of India (SEBI) has introduced the Two Factor Authentication (2FA) model to its stock exchanges by 31st March, 2014. Unlike banking transactions, where anonymity of the user is first preserved, the stock exchange does not provide anonymity and authenticates clients using the User Id and Password combination. Using the User Id and Password combination, an intruder may log in on behalf of a valid customer and manipulate the sensitive transaction, i.e., placing a buy/sell request.

Healthcare: The growth of the digital era has created a new paradigm in the healthcare sector where sensitive patient-related information is being transmitted through different channels across the network. The traditional patient identity authentication mechanism — user id and password combination — is weak and cannot prevent impersonation with unauthorized access. To reduce this risk and prevent manipulative transactions, this hybrid model can be effectively applied, which will add another layer of security. On the other hand, healthcare is one of the prime sectors that takes care of its customers with utmost priority. Thus, the entry barriers need to be maintained very high in terms of security.

Government: Due to digitization, the government is now transmitting different types of sensitive data in real time regarding the citizens, such as a citizen's income details, educational background, criminal record, and biometric recognition. On the other hand, government departments are very keen in protecting such details and do not want them to be misused. To enhance the existing knowledge-based authentication, this hybrid-multifactor model can be employed, which leverages additional competing factors.

Security: The academic institutions have employee databases over a network that contains sensitive employee-related information like the date of birth, address, contact number, health record, and bank account details. The system accepts only the staff member authenticated with the user id/password combination, without considering the intruder's ability to exploit this data and impersonate a valid user to access the network. Thus, it is prone to unauthorized or manipulative transactions. To add another layer of security and reduce the risk of data breach, this hybrid model is implemented.

## 7.1. Banking and Finance

Hybrid Multi-Factor Authentication (MFA) is adopted in the Banking and Finance sector for both persons accessing Internet and Mobile Banking Services. Banks have started biometric authentication using Optical Fingerprint Scanners and Mobile Fingerprint Sensors. Such biometrics are either used alone or with a PIN. A deeper security layer is added by using smart phone sensors along with biometrics. Banks have recently adopted such multi-layer hybrid authentication techniques. This leads to a better customer experience since they do not need to remember any passwords or pins. It also provides security against Man-in-the-Middle Attacks, Spoofing, Phishing, Malware, Replay etc. The hybrid method in this field is cost-effective and easier to manage as only the biometric devices require any additional setup, and existing banking systems can be employed with the analysis [2].

Internet and Mobile Banking services are provided online by commercial banks, development banks, co-operative banks, and financial institutions. These services provide the convenience of transferring money anytime and anywhere. Banks maintain and operate all the functions of a transaction without meeting with the executor, which is dangerous. Data transmission over the Internet can be snooped on, altered, damaged, and have unauthorized access [1]. To fortify the security when a person is accessing Internet Banking Services from a computer, banks have adopted a two-factor authentication process where details are sent over SMS to the registered mobile number of the bank account holder. The person accessing the service needs to fill up these details along with the username and password of the bank account. Apart from the complaints of receiving wrong/dead SMS of one-time passwords from banks, there are serious flaws in these authentications like SIM Cloning, Number Portability, and Man-in-the-Middle Attacks.

### 7.2. Healthcare

Healthcare is a highly sensitive domain, where a small leak can create havoc. Hence, which form of MFA (Multi-Factor Authentication) can provide higher security and scalability? The answer to this considering the rising incidence of cyberattacks and data leaks is hybrid MFA based on two-factor authentication that uses biometric and behavioral analysis. Using biometric characteristics like finger, face, or voice scans, a person can log on to his/her medical information. A sit-stand-sit pattern need not be recorded, but the time between two consecutive sit movements is 0.973 0.55 seconds based on the visual analysis of 8 healthy individuals from 4 different genders. Using this data, each individual must log on to their history every day. If a person sits on the chair for an average of around 19 or more seconds, an unauthorized attempt may be predicted. This second may vary based on the device used for authentication. Hence, in addition to images/finger scans, time-based behavioral analysis may be also considered. In this scenario, there is no need for any additional hardware, and a person must authenticate using their behavior only [1]. Existing biometric data in the current health centers, if not stored in the cloud, can be hacked easily from a particular center; hence, storing only behavioral data in a cloud server can enhance the security and scalability of the system. Cloud computing may provide data access using internet and web services and sceneries to prevent unauthorized access. In this case, the model must consent using a common medium of access for health data across different profiles. Hence, unauthorized access can be traced easily as the peer-to-peer connections among individuals and health care centers can restrict access based on identities. On the other hand, concurrency attacks can be avoided as two distinct individuals can’t be logged into a particular health care center with the same profile, regardless of indirect access [2]. Moreover, medical data are like gold, very sensitive and valuable. Hence, if an individual is diagnosed with any chronic illness or on any medication, he/she can be blackmailed easily. Hence, when the data access should be denied, the platform must consider a certain period of time that is customizable (e.g., 1 hour). In this case, a medical profile is time-sensitive and unauthorized access to its cleaning history cannot be allowed post a particular time period, and if so the abusive act from the healthcare centers can be evicted. To date, the phishing attack, which is directing an individual to an illegitimate page and retrieving input credentials like a username and password, is popular in web service applications such as banking, e-commerce, and social networking. However, as smartphones become ubiquitous in everyday life, the usage of different applications like YouTube, Whats app, Facebook, and banking from mobile wireless devices has increased tremendously aided by the rise of 2G, 3G, 4G, and now 5G. These applications are searching as the new targets and victims of phishing attacks. As a solution, there is a need to deploy mobile device centric phishing detection systems at the application level. Since a particular domain indicates a class of applications having similar functionalities, the used services can be learned separately from the domain perspective and illegal service usages can be identified when targeted against a particular domain. In addition to the growing problem, the notion of social phishing is totally neglected that can retrieve the multifaceted information of an individual, and this model presents a generic framework to detect phishing in the system of systems settings. Besides bank details, blackmail, and hack, controlling data can be used for terrorism. Hence, how to filter harmful access to data is also imperative. With rising incidence of cyberattacks and data leaks, banks like Wells Fargo and companies like E-health have already lost their credibility due to a huge loss of data and finding how much logging is detected or data removed before/after an incident still largely remains unanswered. By developing controlled access to sensitive data in healthcare, government, and defense sector applications can enhance its security.

### 7.3. Government and Security

The government and security sectors have long been reliant on traditional usernames and passwords for systems access. Along with cryptography and the use of firewalls, this is the focus of most effort to help protect sensitive government information. However, these traditional methods are increasingly being bypassed through an increase in the use of sophisticated hacking tools and redirection to social engineering attacks [1]. Consequently, many attempts have been made to bolster security and facilitate authentication using biometrics, which is the use of a person's physical or behavioral characteristics (such as a fingerprint, facial recognition, or a typing pattern) to confirm their identity. Biometrics can be used for verification (in which a person's claimed identity is checked against a biometric template corresponding to that identity) or identification (in which a person's identity is checked against multiple templates in a database) [11]. The combination of biometric authentication systems and behavioral analysis (which is the modeling of a person’s behavior for identification purposes) will be referred to as hybrid MFA systems. These hybrid systems are of high relevance to government and security applications due to the increasing sensitivity of the information the government manages and the rapid progression of hacking and social engineering techniques.

## 8. Case Studies and Implementation Examples

The deployment of Hybrid Multi-Factor Authentication (MFA) using Biometrics and Behavioral Analysis has been applied in various sectors to enhance the security of high-risk transactions. However, due to the innovative nature of this approach, limited information is available regarding its real-world implementations and use in practice.

Case Study: Travel and Authentication via Biometric Facial Recognition. This is an ongoing project in Dubai, UAE. Facial recognition is usually carried out with one image of a person and then matched with an identification source. This approach has limitations due to the failures in the enrollment stage. Similarly, single image gathering and consequent matching with exemplar face-images is usually of high accuracy but can fail if a person is in unique conditions, like a beard or an unexpected expression. For obscured faces, the biometrics cannot be analyzed [5]. As a new approach, the Institute of Advanced Technology and Space Sciences in Dubai is analyzing the same face-video gathered including multiple frames. This reduces the chances of misidentifying a person. Evaluation and subsequent matching between a gallery face-video and a query has been carried out. Face-videos are analyzed using novel descriptor that has been validated in comparative testing with the state-of-the-art descriptors. Long high-speed video sequences have been queried against still images for identification with promising results for obstacles like bad lighting, blurry image, head rotation and so forth. This approach was implemented at Dubai International Airport, allowing travelers to experience a one-stop smart boarding process.

Case Study: Coding of Service Interaction through a Named Entity Recognition Approach for Multi-Proposed Authentication of a Dialog System. Implementing vocal or textual co-evolution, behavioral authentication systems allow for biometric capture in any kind of interaction, favoring context-aware and user-centric systems. Service interactions were coded in a multimodal base of behaviors. The modeling of spoken interaction characters or agents led to categorization of behavioral generalizations. A subsequent recognition experiment with speech-to-text transcription proved statistical significance of verbal account codes applied to authentication of service dialogues. A new methodology of non-verbal coding was prepared and assessed. This work challenges past representative approaches on various fronts [3].

## 9. Future Research Directions

The discussion is focused on potential future research directions and advancements in the field of Hybrid Multi-Factor Authentication (MFA) using biometrics and behavioral analysis. There is always a need for advanced security measures as online systems and devices are increased. Most of the next generation technologies such as automatic authentication and cyber security could be advanced and supported by physical and behavioral biometrics. The advancement of technologies such as artificial intelligence may also play a key role in the evolution of hybrid MFA. The use of artificial intelligence in hybrid MFA applications will play a significant role in decreasing false rejection rate and false acceptance rate. Overall, hybrid MFA technologies should be trying to keep pace with the changing environment of the need for enhanced security in both professional as well as personal devices [7].

MFA technologies should be exploring the potential of the use of number of behavioral biometrics either independently or in concurrence with other physiological biometrics. Recent studies have been using behavioral biometrics such as the use of mouse or touchpad movement patterns and keystroke dynamics patterns independently However, these patterns were either not combined with other behavioral biometric parameters or were not studied in concurrence with other physiological biometrics. For continuous authentication either keystroke dynamics or mouse movements were mainly used and no applications have been found which uses the combination of these biometrics to authenticate users continuously. More research is needed to understand the potential applicability of these life indices biometric parameters with other physiological biometric parameters [3].

## 10. Conclusion and Summary

A comprehensive conclusion and summary of the key findings and insights presented throughout the paper is provided. The significance and potential impact of the paper's discussion on hybrid Multi-Factor Authentication (MFA) using biometrics and behavioral analysis on diverse sectors is highlighted. Aspects are recapitulated to strengthen the considerations' importance. The variety of hybrid MFA systems has been sufficiently analyzed such that different approaches and the requirements and methods for the implementation of a specific approach have been discussed and understood. A hybrid MFA system using voice recognition complemented by behavioral analysis has been presented, and possible actions for the research and development of the completion of such a system have been suggested. As technology continues to advance, security measures become more sophisticated and provide heightened protection against new vulnerabilities. However, usage of such high-security requirements represents greater challenges to individuals, as required steps to fulfill basic actions become more daunting with the introduction of more complex procedures. Therefore, consideration of such issues in future developments of MFA systems should not be overlooked. The use of a hybrid MFA system would be beneficial across a variety of sectors, with such systems capable of protecting banks and government institutions while providing individuals enhanced security with personal banking and identification services. [12] Successful initial implementation of one approach, such as the voice recognition MFA, might spur the additional avenues of research suggested earlier and would advance further development of such systems across multiple applications. On the other hand, such positive outcomes should not be taken for granted, as Hawala is a financial transaction based network which exists and thrives in parallel with standard banks and global financial institutions; in danger of being far more sophisticated and untraceable such unregulated implementations may become further complements or replacements of parallel forms of governmental identification systems as mobile phones become ubiquitous and biometric identifiers evolve into universal identification codes able to link everything.

## 11. References:

[1] K. Modi and L. Devaraj, "Advancements in Biometric Technology with Artificial Intelligence," 2022. [PDF]

[2] C. S. Koong, T. I. Yang, and C. C. Tseng, "A User Authentication Scheme Using Physiological and Behavioral Biometrics for Multitouch Devices," 2014. ncbi.nlm.nih.gov

[3] L. Filipe Machado Malhadas, "Perceiving is Believing. Authentication with Behavioural and Cognitive Factors," 2017. [PDF]

[4] T. Saad Mohamed, "Security of Multifactor Authentication Model to Improve Authentication Systems," 2014. [PDF]

[5] A. Verma, V. Moghaddam, and A. Anwar, "Data-driven behavioural biometrics for continuous and adaptive user verification using Smartphone and Smartwatch," 2021. [PDF]

[6] N. Damer, "Application-driven Advances in Multi-biometric Fusion," 2018. [PDF]

[7] H. Zhang, D. Singh, and X. Li, "Augmenting Authentication with Context-Specific Behavioral Biometrics," 2019. [PDF]

[8] M. Abuhamad, A. Abusnaina, D. H. Nyang, and D. Mohaisen, "Sensor-based Continuous Authentication of Smartphones' Users Using Behavioral Biometrics: A Contemporary Survey," 2020. [PDF]

[9] M. Singh, R. Singh, and A. Ross, "A Comprehensive Overview of Biometric Fusion," 2019. [PDF]

[10] R. Dave, N. Seliya, L. Pryor, M. Vanamala et al., "Hold On and Swipe: A Touch-Movement Based Continuous Authentication Schema based on Machine Learning," 2022. [PDF]

[11] B. Alharbi and H. S. Alshanbari, "Face-voice based multimodal biometric authentication system via FaceNet and GMM," 2023. ncbi.nlm.nih.gov

[12] J. Tian, "Authenticating Users with 3D Passwords Captured by Motion Sensors," 2018. [PDF]

[1] TUG, Institutional members of the TEX users group, 2017. URL: http://www.tug.org/instmem.html .

[2] R Core Team, R: A language and environment for statistical computing, 2019. URL: https://www.R-project.org.

[3] S. Anzaroot, A. McCallum, UMass citation field extraction dataset, 2013. URL: http://www.iesl.cs.umass.edu/data/data-umasscitationfield.

# Hybrid Secure Routing in Mobile Ad-hoc Networks (MANETSs)

BOUFAIDA SOUNDES OUMAIMA[1], BENMACHICHE ABDEMADJID[2] and MAATALLAH MAJDA[3]

[1] *Faculty of Technology,Departement of Computer Science Universty of Chadli Bendjedid El Tarf, Algeria*

*s.boufaida@univ-eltarf.dz*

[2] *Faculty of Technology,Departement of Computer Science Universty of Chadli Bendjedid El Tarf, Algeria*

*benmachiche-abdelmadjid@univ-eltarf.dz*

[3] *Faculty of Technology,Departement of Computer Science Universty of Chadli Bendjedid El Tarf, Algeria*

*maatallah-majda@univ-eltarf.dz*

## Abstract

Because wireless communication is dynamic and has inherent defects, routing algorithms are crucial in the quickly evolving field of mobile ad hoc networks, or MANETs This study looks at the many security problems that MANETs encounter. These problems, which pose major risks to network performance, include flooding, sinkhole, and black hole assaults To address these challenges, we introduce the Hybrid Secure Routing Protocol (HSRP), which enhances the security and robustness of routing operations by fusing trust-based tactics with cryptographic approaches. HSRP combines the strengths of both proactive and reactive routing strategies, enabling it to adapt dynamically to evolving network conditions while protecting against malicious activities. We use extensive simulations with Network Simulator (NS-2) and a thorough review of the literature to assess HSRP's performance under different attack scenarios. The results show that, in comparison to traditional protocols, HSRP increases throughput and decreases latency, hence improving routing efficiency while simultaneously bolstering data transfer security. With uses in vital domains including military operations and disaster response, this study provides a scalable and workable approach for safe routing in MANETs. The findings highlight how crucial it is to include cutting-edge security features in routing protocol design in order to guarantee the dependability and integrity of MANETs in practical situations.

## Keywords

Mobile Ad-Hoc Networks (MANETs), Secure Routing Protocols, Hybrid Routing, Trust Management, Cryptographic Techniques, Attack Mitigation, Dynamic Topology

## 1. Introduction

Mobile Adhoc Networks (MANETs) are characterized by self-organization, without the use of any centralized control or administration. As these networks are infrastructure less, the nodes in the network act as routers. The routes in the network are discovered by the nodes using different routing protocols. MANET protocol diversity, and user mobility, are responsible for the variation in the nature of security attacks on routing protocols. Routing protocols of MANETs lack security features. An insecure routing protocol can wreak havoc on network security, making it easy for adversaries to exploit it for malicious purposes while remaining undetected. Sometimes these protocols start misbehaving because of security attacks and distort various services of the environment. Security problems associated with MANETs and solutions to achieve more reliable routing are the focus of these studies [1].

A distributed denial of service (DDos) attack is a severe threat. Similar to denial-of-service attacks on local area networks, internet application servers can also be attacked. MANET can be corrupted by overwhelming radio frequency (RF) signal. Nodes in the vicinity of such attack would be unable to communicate. Much larger MANETs would be difficult to corrupt with RF jamming signal, but smaller MANETs can be congested very easily [2]. In the case of a jamming attack, nodes cannot communicate due to RF signal interference. The network can be most

[*] Corresponding author.

[†] These authors contributed equally.

aleksandr.ometov@tuni.fi (A. Ometov); t.princesales@utwente.nl (T. P. Sales); manfred.jeusfeld@acm.org (M. Jeusfeld)

0000-0003-3412-1639 (A. Ometov); 0000-0002-5385-5761 (T. P. Sales); 0000-0002-9421-8566 (M. Jeusfeld) vulnerable if only network control packets are jammed as nodes are most susceptible to this.

### 1.1. Background of Mobile Ad-Hoc Networks (MANETs)

A Mobile Ad-hoc Network (MANET) consist of mobile nodes forming a mesh network without fixed base stations. Generally, Mobile Ad-Hoc Networks (MANETs) are infrastructure less, self-configured, multi-hop wireless mobile networks formed in dynamic topology with mobile nodes. Depending on the applications, size, etc. some nodes can run on batteries, which limits the energy at respective nodes [1]. Speedy deployment at any desired location is an advantage for MANET that is enabled by portable laptops, handheld devices, etc. setting-up a network without the need for infrastructure is the goal of such networks. Sending and receiving of information can take place in MANETS, with packets routed by intermediate nodes that replicate the functionality of routers in a wireless network by forwarding packets based on information gathered about transmitter nodes. In wireless networks, radio wave transmission takes place for communication. Radio transmission has properties like reflection, diffraction, delay spread, and scattering that lead to problems of multipath propagation. High-frequency smaller wavelength radio waves in an attribute frequency range of AKHz to GKHz are also heard among the MANET problems where pulses generated by antenna becomes wide and increase bit error rate.

Wireless mobile networks have recently gained significant interest due to the pervasive use of wireless technologies, especially wireless LAN’s [3]. Wireless technologies enable economic and easy installation and are suitable for mobile users. Wireless mobile networks can be broadly classified into two categories: infrastructure based wireless networks and mobile ad hoc networks (MANETS). A wireless communication network with a base station called access point or land station, is termed as infrastructure based wireless network. However, for an ad hoc network no predefined fixed infrastructure exists. A mobile ad hoc network is a collection of autonomous nodes which can move in and out of the network freely. Hence, the network topology changes dynamically in a non-predictable manner. In addition to these complexities, wireless communication is susceptible to noise, interference, multipath propagation, fading, etc. All these factors make the design of MANETs challenging.

### 1.2. Importance of Secure Routing in MANETs

In the last decade, a great deal of research has been conducted on wireless networks, inspired in part of their potential applications. Wireless networks are networks that interconnect mobile devices without a wiring infrastructure. Current standards support devices with short range of communication from 10 m up to 100 m, such as Bluetooth and IEEE 802.11. MANET needs a security solution to offer protection against various types of attacks. In MANET, a mobile node(s) can join or leave the network at any time without any prior notice which makes a consistent and reliable routing very difficult. These networks are more vulnerable to various types of attacks because mobile nodes can work on the network with less physical protection. Mobile Ad-hoc Networks (MANETs) consists of group of mobile nodes that communicate with each other without any pre-existing fixed infrastructure. A MANET is an autonomous system of mobile hosts connected by wireless links. Each node in MANET acts as both routers and hosts which can move in arbitrary fashion. But due to such nature of networks, one hop or multi-hop wireless communication among the nodes is exposed to various security challenges and attacks involving different threats and at different scales. Either by machinery or human operators, in any situation (natural disasters, emergency operations, war fields), adversary users can attack to all or one selected node(s) by hindrance or by benign nodes to capture and discloses some private information.

Routing protocols are necessary to discover paths through an ad-hoc network between pairs of wireless devices. The alternative paths are less loaded, or higher data rate. A proper routing protocol selects between several paths for which the connections were originally set up. However, to meet the actors' requirements, the protocol must also keep the paths selected up to date. With a continually changing network topology, errors can occur in the paths. For detecting networking failures there is generally periodic monitoring of the paths as well. Due to hostile conditions, the nodes in MANET face attacks that can do the routing information modification or dropping of any messages. MANET provides perfect environment for intruder, malicious user which can corrupt the entire routing table due to the absence of base station. As this network does not have any centralized or fixed infrastructure, a mobile node(s) can join or leave the network at any time which may cause hindrance in its connectivity. So there is a need of secure routing protocol that can offer reliable and stable routing.

### 1.3. Overview of Trust-Based and Cryptographic Approaches

A routing protocol can be defined as a set of rules by which a node/host obtains a route to all other nodes within the same logical network. There are different protocols for different types of networks i.e. wired, wireless, ad-hoc and peer-to-peer networks. Routing protocols for Wireless network: AODV, DSR, SRP, DSDV, WRP. Trust Issues in Routing is an effort to study the security problems associated with MANETS and the methods or solutions through which routing can be made more reliable [1]. There is a certain class of issues that occur due to Ad-hoc Network Characteristics.

Trust-based approaches use social concepts like reputation to make a contribution toward securing networks. Reputation systems use past recommendations or individual observation and assess trust bounds of nodes based on those observed interactions and/or recommendations, to either allow or deny forwarding packets via nodes. Cryptographic approaches leverage over-sharing trusted authorities and public keys to share security parameters, which are subsequently used to create security associations. Given that such assumptions are not viable in MANETs, cryptographic techniques mainly rely on digital signatures or certificates to provide security against tampering and/or forgery and replay and/ or masquerading attacks. Unfortunately, such techniques are susceptible to internal attacks by compromised nodes that violate terms of use of the utility arguably for individual pecuniary profit. External attacks are also possible by attempts to forge signatures or assimilate false certificates therefore either being able to fabricate messages or being tamable to misuse [4].

## 2. Trust-Based Routing in MANETs

A Mobile Ad-Hoc Network (MANET) is a self-configuring Infrastructure less network of mobile devices connected by wireless links. Each node in the network may act as a host and as a router. It works on cooperative basis to perform network functions, thus networking protocols designed for wired networks are not directly suitable for MANETs. As the nodes are mobile, they are free to move in any direction and there is a frequent change in topology. These constant changes create challenges such as availability of complete and correct node information in a timely manner. Trust, security and quality of service are essential criteria in selecting a route to other nodes for transferring data among them [1]. This project studies different trust-based routing protocols in MANETs. A well analyzed trust model is implemented in preventive as well as reactive approach to enhance the performance of networks.

A Mobile Ad-hoc Network (MANET) is a self-configuring infrastructure less network of mobile devices connected by wireless links. A mobile ad-hoc network (MANET) is a self-configuring infrastructure less network of mobile nodes connected by wireless links and having dynamic topology. Each node in the network may act as a host and a router. It works on cooperative basis to perform network functions; thus networking protocols designed for wired networks are not directly suitable for MANETs. The nodes can come together, communicate within each other and leave anywhere anytime depending on their mobility. As the nodes are mobile, they are free to move in any direction. There is a frequent change in topology. Hence the nodes may enter and leave the network at random times. These constant changes create challenges such as availability of complete and correct node information in a timely manner [5].

### 2.1. Concepts and Principles

With the advancement of computing and mobile devices, wireless ad hoc networks, or more precisely, Mobile Ad Hoc Networks (MANETs), have become an area of tremendous interest and research. A MANET is a collection of mobile nodes which communicate with each other. These nodes are not fixed and can move while transmitting data packets. Nodes are connected to each other dynamically in an arbitrary fashion, but communication takes place over multiple hops due to the limited transmission range of individual nodes [1]. Nodes can be added to or deleted from the network at any time, and the network topology keeps changing randomly and frequently. These types of networks have become essential for various military applications, such as battlefield surveillance, combat operations, search-and-rescue operations, etc. Moreover, as each node acts as both a host and a router, there is no need for prior installation of base stations or centralized administration. Furthermore, since all the nodes are free to enter and leave, there is no fixed topology which results in the dynamic and infrastructure-less nature of a mobile ad hoc network and a pose for secure routing protocols and the development of active security [5].

Secure routing Protocols are essentially responsibility for connecting mobile nodes in a mobile ad hoc network (MANET) securely. Mostly nodes in MANET are mobile and each node acts as both a host and a router. Furthermore, there is no fixed topology and all nodes are free to enter and leave. So nodes are more introduced to the vulnerability of being misused by attacks. All routing packets are forwarded between nodes in a free manner. This creates a chance to perform various security attacks during routing. The Secure/Trusted Routing algorithm will try to increase the security of routing in MANETS. The purpose of daily computing is to securely connect all users in an unknown network. In the context of mobile ad hoc networks (MANETs) and peer-to-peer systems, the users are mobile devices that dynamically join and leave the network and act as servers and clients. Meanwhile, these devices cannot be fully trusted and can act cooperatively or maliciously.

### 2.2. Trust Metrics and Models

In trust-based routing, nodes should evaluate the level of trust to communicate with neighboring links. The most common approaches to evaluate trust metrics are quantitative metrics and qualitative metrics [6]. In mobile ad-hoc network (MANET), links between nodes are created dynamically; users should build a trustable relationship before establishing a communication link. A quantitative trust metric is defined by a node that uses communication requests and responses to calculate the degree of trust in the connection link. Trusting nodes are classified as good, neutral, and bad based on numeric values, where 0 <= trust < 0.5 is bad, 0.5 <= trust < 0.8 is neutral, and 0.8 <= trust <= 1 has a good trust level. The trust level of a communication link is calculated using the number of direct transactions of one node and its neighbors (engagement process), assessments provided by the other node (reputation process), and observations made by the intermediary nodes (recommendation process). On the other side, each node maintains a trust table that keeps track of how other nodes are assessed with reference to the communication transactions in which it was part either as a source node or as a destination node.

A positive or negative weight is assigned based on their behavior. On the other side, the qualitative metrics may be proposed as a node assigns a qualitative label to the communication links: link strong if no bypass is detected, link normal if there are some bypass but are less than a threshold, and link weak if there are many bypasses. With qualitative trust metrics, it can be easier to justify a certain behavior to outside observers because it is less sensitive to small variations [7].

## 3. Cryptographic Techniques in MANETs

### 3.1. Digital Signatures

Digital signatures are based on asymmetric key cryptography, which uses two keys: the private key for signing and the public key for verification. Using digital signatures, the sender can generate a signature for the data using its private key. The receiver can verify the authenticity of the data and signature using the sender's public key.

Digital signatures have become a fundamental tool in achieving security properties such as data authenticity and non-repudiation. However, due to resource constraints and special characteristics, the implementation of digital signatures in MANETs needs to be adapted. To withstand forgery attacks on mobile nodes and prevent malicious activities such as altering the source of a message or replaying an old message, digital signatures need to be personalized for each node in a MANET [1]. Thus, asymmetric key cryptographic parameters for digital signatures are generated apart from the regular parameters of a mobile node. Long static secret keys (private keys) are safely stored in a tamper-resistant smart card. The public keys are stored in a public directory and maintained by a trusted key generation center.

### 3.2. Encryption Algorithms

In the next section, Mobile Ad-Hoc Networks (MANETs) are described, where a proactive encryption algorithm is suggested. There are two networks involved: the first is the infrastructure network, and the second is the anehdrouter network with nodes on the roadside using MANET and DSRC protocols. In the anehdrouter network, each node is the router. Mobile devices use the device's Wi-Fi or Bluetooth to connect to a node. Then the node chooses a path and routes packets until it reaches the destination node. Therefore, there is a possibility of being exposed to external attacks. Tiny encryption algorithm (TEA) is proposed to encrypt data. TEA belongs to the class of block ciphers with a block size of 64 bits, utilizing a 128-bit key with a simple structure. It consists of 32 rounds of processing where one round of TEA accepts a plaintext consisting of two 32-bit halves. A compression function that works iteratively applies basic operations, including addition, XOR, and shifts. The design of TEA is a trade-off between performance and resistance to cryptanalysis. TEA was initially designed for 32-bit processors, although more hardware-efficient implementations can be easily adapted. By using a large block cipher such as the AES, the Man-in-the-middle will be more expensive [1].

Then another block cipher is suggested which known as twofish algorithm. Two fish was selected as one of five finalists for the advanced encryption standard (AES). It has a block size of 128 bits and a key size of up to 256 bits. It has some attractive features such as: Two fish is faster than triple DES on most platforms. It is believed to be the most efficient AES finalist for software implementations on the vast majority of computing platforms. Claiming that it has a good performance in hardware implementations as well. It is known to be very secure to all known attacks, and the design methodology of two fish was specifically intended to be practically attack resistant. No attacks against two fish more efficient than exhaustive key search are known. Two fish uses the same core structure as blowfish: a 16-round Feistel network using two independent subkeys for each round. The structure of two fish is very different from that of family of block ciphers, including DES, AES and RC529 which rely entirely on lookup tables for diffusion.

## 4. Challenges in Secure Routing in MANETs

The mobile ad-hoc networks (MANETs) are wireless networks distinguished by the lack of fixed infrastructure and a dynamically changing topology where each mobile node acts both as a host and a router without central administration, thus forwarding traffic for other nodes. MANETs permit nodes to move freely and can quickly perform communication in the absence of fixed infrastructure. These networks find many applications in rescue operations, military deployments, commercial product campaigns, and environments with no fixed infrastructure like meetings, conferences, and classrooms [9]. Each node in a MANET employs wireless communications to reach nodes within its transmission range, while other nodes outside this range receive packets after it has been relayed by other nodes – thus creating a multi-hop network. However, wireless channels are inherently error-prone because of various reasons like transmission noise, interference, mobility of nodes, and signal distortion. Topological changes take place because of node initialization or expiration, node mobility, link failures, or interference created by obstacles, causing routes to be free or unestablished. Due to this, the structure of the network continuously generates routes discovery and maintenance traffic, thus generating high overhead and processing delays [1]. Security is an essential constraint due to the increase in the number of applications for mobile ad-hoc networks. MANETs are prone to security threats as the medium is open, and therefore sensitive information can be compromised. Furthermore, the protocols, information, and bandwidth are open to outside access, where malicious users can join the network to gain confidential information or affect the routing, thus draining the resources. Thus with a highly dynamic network topology combined with resource constraints (limited battery power, on-board memory, CPU power, information processing/forwarding capacity, and collision susceptibility) represent a major challenge for the design and implementation of secure routing protocols.

### 4.1. Dynamic Network Topology

The mobile nodes in MANETs compose the network's dynamic topology through the movements of nodes, which leads to changes in links over time. Several factors contribute to the dynamic topology in a MANET, including geographic location, node mobility variation, network environment, and life period. Each factor is elaborated upon here.

Geographic Location: The geographic location of a node affects its movement. For instance, in a battlefield MANET, nodes comprising mechanisms like cameras, radars, and air-attack systems located nearby each other (i.e., close in their geographic location) are likely to move in similar geographic terrain and thereby lead to similar movements. As a result, these nodes form a dynamic topology.

Node Mobility Variation: The variation in node mobility influences dynamic topology. The MANET scenarios having low mobility or high mobility create similar problems in maintaining the topology. For instance, in a network with low node mobility, the routes are stable and remain intact, while nodes move from one segment to another rapidly; often, rerouting does not occur within limits of the time-to-live (TTL) that specifies the packet lifetime.

Network Environment: Network environment determines the node movement pattern. For instance, in a scenario of rural/guided movement, nodes travel in a uniquely directed path and access the network subsequently or disband it (i.e., there is no further opportunity for communication). This leads to formation of two topologies: (1) initial topology prior to movement, and (2) final topology after movement.

Life Period: The life period of all nodes determines the node movement pattern. For instance, when all the nodes are equally likely to die and thereby form a dynamic topology and network segment, there is an abrupt increase in disconnected nodes. In addition, in a scenario where all nodes are assigned the lifetime (TTL) such that they randomly die after the period, there is an abrupt shift from connected to disconnected topology.

### 4.2. Resource Constraints

The design and deployment of routing protocols in MANETs are complicated by resource constraints. Each node in the MANET works using resources like energy, bandwidth, and processing power. The nodes are battery powered, resulting in limited energy resources that directly affect the lifetime of the network. In a MANET, every node has a wireless link of a limited bandwidth, which when not utilized correctly may cause congestion and packet dropping in the transmission. The resource constraints are not limited to energy and bandwidth, there are several protocols like IPsec that require high nodes processing capabilities that many of the mobile nodes cannot meet [1].

Considering the resource constraints, the secure routing protocols must be designed, which should consume less energy and bandwidth in the secure route establishment and maintenance. The protocols not only have to look for the most energy-efficient route, and less delay route, but they also have to look for routes with less overhead packet size for its security schemes.

## 5. Hybrid Secure Routing Protocols

Hybrid approaches that take advantage of both trust-based and cryptographic approaches provide the best of both worlds: the flexibility, scalability, fault-tolerance, and low overhead of the trust-based approach coupled with the security assurances provided by public key infrastructures and cryptographic mechanisms. Recent works in this area are completely classified, compared, and discussed in detail. The trust concept is based on subjective evaluation of other agents and is getting rapidly popular because it provides a decentralized way to ensure security where the existing solutions fall short [5]. However, trust-based solutions cannot completely substitute existing security mechanisms because of their limitations. On the one hand, the cryptographic mechanisms provide a high level of security with strong characters; but, on the other hand, they are too rigid, inflexible, and costly for many scenarios. During certain phases of trust evaluation, such as in the proactive stage, these solutions require costly signature verification for all received messages which, on very low power nodes, might be a problem. To the authors' knowledge, there is the initial work in the area of hybrid secure routing protocols for ad hoc networks. Numerous trust-based protocols have been proposed in the last years; most of them rely on accurate modeling of the network to assess node trust values [1]. Although accurate, trust models highly depend on particular scenarios for which they were designed. When these models are used in different scenarios or when the modeled systems change, the correctness of the model will decrease and it might provide completely wrong predictions or estimations. All these predictions or estimations should be handled with care; thus, trust and trust-based systems should be considered as crude methods that have only relative and subjective effectiveness. For example, in extreme situations, such as islanded systems (i.e., where the network nodes are temporarily separated from the global networks or from the main trusted nodes which are used for trust evaluation), inaccurate predictions are possible. In alternative scenarios such as complete cooperation among nodes, all those assumptions become invalid, and under such a cooperation assumption, nodes trust each other completely and there is no need for trust management. Further, in trust-based solutions, the trust values reflect the past behavior of nodes, which might be not appropriate for the current state (e.g., during the initial phase in which the view on the network is incomplete).

### 5.1. Integration of Trust-Based and Cryptographic Approaches

The integration of trust-based and cryptographic approaches within hybrid secure routing protocols is examined. The regulatory standards and governance structures help to ensure cooperation and reduce the threat of attacks. The work in the trust approach gives more confidence in a route, thus, cracking down the information pass-through easily. In contrast, the cryptographic approach deals with the security margins and enhances the level of encryption. The routing protocols under the existing hybrid approach balance the trust and cryptographic schemes. Hence, the hybrids work in complementary nature of approaches to fortify the security in MANETs [10].

The collaboration of cryptographic and trust-based approaches gives not only the strengths but also reduces the drawbacks, as shown in Table 6. The cryptographic approach helps in the exchange of secret keys and thus eliminates the trust creation in each communication. The advantage of these hybrid protocols is to lessen the penalty cost during the time of attacked nodes which do not cooperate and misbehave. Moreover, the prevention of attacks is handled at the routing community level, i.e., the detection of misbehaving nodes during the route construction phase itself. Thus, the likelihood of such nodes coming into the communication area is minimized [11].

### 5.2. Advantages of Hybrid Protocols

Hybrid secure routing protocols exploit the advantages of both trust-based secure and cryptographic secure routing protocols. The trust-based secure routing protocols are itself a good option for secure routing under partial high threats as they are well suited for such environments and gives high performance. The cryptographic secure routing protocols, on the other hand, are better option in case of high threat environment. However, cryptographic secure protocols still have some robustness. Therefore, the combination of trust-based secure and cryptographic secure routing protocols will offer better performance as compared to trust-based secure routing protocols in high security threat environment and improve robustness as compared to cryptographic secure routing protocols in moderate security threat environment [10].

In any given environment, either the nodes will have the partial trust over the nodes or the nodes will not trust each other at all. The trust-based secure routing protocols are an appropriate option for the environment when either the nodes will have trust on some nodes and those nodes are suspected to be malicious and at the same time it is needed to secure the routing protocol from them. While the cryptographic secure protocols offer higher security features, but they are less resilient and adaptive as compared to trust-based protocols. Since both the techniques offer significant advantages when used in their appropriate environments, the combination of trust-based and cryptographic secure routing will produce a hybrid secure routing protocol which will offer good resilience and adaptiveness due to the trust-based protocols and at the same time will have good security features due to the cryptographic secure routing protocols.

## 6. Research Methodology

### 6.1. Literature Review

The literature review was conducted in order to examine existing knowledge and research on mobile ad-hoc networks (MANETs) security and hybrid routing protocols. A search of scholarly articles and conference papers was carried out using Google Scholar, Springer Link, IEEE Xplore, and Wiley Online Library. The following keywords were used: “Mobile ad-hoc networks,” “MANETS security,” “Security in MANETS,” “Hybrid routing protocols,” “Hybrid secure routing in MANETS,” and “MANETS routing with security”. Articles published between 2005 and 2020 were examined, as studies conducted after 2020 are currently limited to pseudocode or simulation design only. As a result of the search, a total of 27 scholarly articles and conference papers were selected for full-text reading based on their title and abstract. This process resulted in 11 scholarly articles and conference papers that provided detailed information on the research topic and were selected for this literature review.

MANET is characterized by mobile nodes, multi-hop wireless connectivity, infrastructure less environment and dynamic topology. A recent trend in ad-hoc network routing is the reactive on-demand philosophy where routes are established only when required. Stable routing, security and power efficiency are the major concerns in this field. This paper is an effort to study security problems associated with MANETS and solutions to achieve more reliable routing. To understand the attacks in an efficient way related work has been categorized into active and passive attacks, which helps us in developing counter action for the work [1]. The ad hoc environment is accessible to both legitimate network users and malicious attackers. Legitimate users can access the network only when routing protocol is employed. Malicious users, having knowledge of routing protocols are able to join the network without any authentication, which results in the generation of false routing information. False sequencing, hijacking and eavesdropping are some of the examples of attacks. The study will help in making protocol more robust against attacks to achieve stable routing in the routing protocols [2].

### 6.2. Data Collection and Analysis

The protocol evaluation process begins with data collection. Four different simulation scenarios were built around the same mobile ad-hoc network with 50 nodes, node speeds between 0-10 m/s, and a network area of 700m × 700m. Then, Hall of Fame, Message Dropped, RREQ and RREP packets generated in each scenario were logged. COMNET II was used to run simulations of all scenarios. COMNET II is an object-oriented simulation language with a graphical user interface suitable for modeling various phenomena [12]. It includes building and troubleshooting simulation models, running simulations, collecting and analyzing output data, and generating animations and reports. Scenarios 1 and 2 represent no attack condition, while 3 to 6 represent black hole attacker with two protocols.

The evaluation metrics include Hall of Fame Packets Count (HF), Message Dropped Count (MD), RREQ Packets Count (RREQ) and RREP Packets Count (RREP). Hall of Fame is a collection of packets that each node stores and compares with its own. The score is incremented by one each time a packet is stored by the node. It discards packets if it has already received and stored the same one. The HF count is used to evaluate how well the protocol propagates information throughout the network. The more efficient it is, the higher the HF count. In a badly designed protocol, the network may experience broadcast storm problem, resulting in low HF count. Message Dropped Count is the total number of data packets that the destination node called in to be sent but were never received by the MAC layer. These packets are counted only if the forwarding node has delivered the final control packet to the MAC layer. If at least one such same control packet was successfully delivered to the MAC layer, these packets won't be counted as dropped. In no attack scenario, the MD count is expected to be low [13]. If the count is unusually high, it indicates that the protocol is not functioning properly.

# 7. Case Studies and Experiments

[13] There are two case studies done on MANETs. The first is 'Proposed Scheme for Secured Routing in MANET' and the author is Nidhi Goyal and Sushil Kumar. This paper presents an effort to study the security problems associated with MANETS. It also gives solutions to achieve more reliable routing in MANETs through the improvised threshold scheme. MANETs are characterized by mobile nodes, multihop wireless connectivity, infrastructure less environment, and dynamic topology. Stable Routing, Security, and Power efficiency are the major concerns in this field [1].

The research work carries out a study of various routing protocols on the basis of performance in terms of throughputs with respect to 50 mobile nodes and distances of 1000, 1500, and 2000m. The set of standard protocols is bandwidth, flow, and size. A Mobile Adhoc Network (MANET) is a collection of mobile nodes forming a temporary network without the use of any fixed backbone network. It has the potential to be applied in areas such as disaster relief operations, military applications, and hospital environments.

## 7.1. Simulation Environments

The network simulator (NS-2) is a commonly utilized simulation language for packet-level simulation of different wireless network protocols. NS-2 provides an object-oriented simulator with support of C++ and OTcl simulation languages. NS-2 aims to exhibit the modeling of protocols and to run simulation scripts needed for analysis and visualizing simulation results [14]. For performance comparison of routing, the path with source and destination nodes during simulation of each run is also depicted in NS-2 as a text file. The data Mobile Ad Hoc Network is implemented with independent node distribution.

The number of mobile nodes is taken in the range of (20, 50, 100), and Maximum Speed is taken in the range of (5, 10, 15). The packet size is considered as 512 bytes, and data rate is taken as 2.5. The routing protocols AODV and DSR are compared under all conditions, and graphs involving throughput, jitter and end-to-end delay are plotted in MATLAB for analysis of results. The network area is fixed with 1000m x 1000m for all conditions [1].

## 7.2. Performance Metrics

To evaluate the efficacy and efficiency of secure routing protocols modeled for MANETs, performance metrics have been employed. They serve as quantitative measures in determining the quality of routing protocol mechanisms. A comparison of performance metrics yields a basis for a comprehensive understanding of a routing protocol's ability to route packets within a network. The ideal characteristics of a performance metric may include observability and reproducibility [14]. The observability of a performance metric indicates that its measurement can be interpreted, implemented and monitored in all future instances of case studies, while reproducibility indicates that the outcomes of an experiment are comparable to other conductors and can be replicated with consistent results. Between the different performance metrics employed to evaluate routing protocols in MANETs, at least one metric should utilize an analysis of the network's bandwidth, one should employ mechanisms of end-to-end delay stability, one should analyze packet delivery fractions and one should consider the overhead ratio of control packets to data packets [15]. All parameters considered in the performance evaluation whose inclusion is determined as necessary for complete and thorough analysis, based on industry and scientific standards. Through simulations conducting a number of scenarios including a variety of node densities and speeds, results are obtained that reflect upon the performance of the routing protocols as a whole. The outcomes permit a comparison and analysis of the results of each scenario, demonstrating trends in performance across different environments and establishing a foundation for conclusions regarding the routing protocols. Such conclusions are especially relevant to the implementation of a routing protocol model in an environment such as an emergency response, where the performance of routing protocols could have both a critical and significant impact.

# 8. Results and Findings

The results obtained from the case studies and experimental analysis are detailed in this section. Apart from graphs and statistical data, a comparative analysis with a set of existing protocols is given. These results are categorized and presented on the basis of the significant functions of the proposed hybrid protocols. With these protocols, defense and detection mechanisms to combat DOS attacks, stealthy attacks, and the trade-off between these two aspects are highlighted and graphically represented. In addition, the efficiency aspect is analyzed to substantiate that these hybrid protocols are better in terms of performance than the existing protocols. The proposed hybrid set of protocols also includes HSRDP and MSRDP, which serve the purpose of defense mechanisms for DOS and stealthy attacks, respectively. Along with the robustness of the individual protocols in mitigating specific attacks, a hybrid application of these protocols is also analyzed to study its dual combat against both DOS & stealthy attacks. For this, a model is built with the hybrid application at the AODV layer. In Layer 3 (Network layer – here it is AODV), HSRDP, MSRDP, and Hybrid protocols are integrated, while RIP, DNS, and DSRp remain at their original state. The results from such a hybrid implementation, which combine both protocols at once, are discussed in detail. Apart from hybrid analysis at the AODV layer, standalone core implementations are also executed similar to the earlier graphs to study individual performances. To validate the implementation set up, network performance is analyzed with AODV in its default state (without hybrid defense). The performances are graphically displayed to indicate the impact of stealthy and DOS attacks on routing metrics.

### 8.1. Comparison with Existing Protocols

In consideration of the vulnerabilities associated with both proactive and reactive routing schemes for mobile ad hoc networks (MANETs), a revised version of the hybrid secure routing protocol previously proposed has been developed. The revised routing protocol has been simulated in connection with a complete version of the secure link state routing protocol (SLSP), which was found to be efficient and secure during preliminary simulations. The objective of the hybrid secure routing scheme is to emulate the characteristics of proactive and reactive protocols, providing a novel option for MANET routing that improves the security of the network. The revised protocols have been subjected to full simulations to assess protocol efficacy and reliability under both typical network scenarios and extreme cases [2].

A comparative assessment is conducted of the performance and security attributes of the new hybrid secure protocols against existing routing protocols for MANETs. Observations and conclusions are offered. Routing protocols for MANETs have been analyzed in terms of security characteristics and simulation performance in relation to randomized core group attacks on MANET systems utilizing different types of proactive, reactive or hybrid routing. A proactive protocol routing protocol was found to be most affected by such attacks. Of the other routing protocols investigated, the random wait time optimization procedures were found to be the most effective in terms of performance and effectiveness in circumventing core group attacks [16].

### 8.2. Security and Efficiency Evaluation

The evaluation of proposed hybrid protocols is conducted via extensive simulations using the Network Simulator 2 (ns-2) considering the parameters of packet delivery ratio, end-to-end delay, routing load and throughput. The performance of the proposed protocols, which provide β-bandwidth routing as well as security, is compared with existing protocols that only provide attack-free routing and β-bandwidth routing. The attack used here is reduction of the maxima value in the route request packets. The σ-Max routes with benign nodes only are spoofed with malicious nodes (attack) using which the performance of the protocols is evaluated.

The performance of the proposed hybrid protocols is evaluated in terms of ω and ψ parameters. To evaluate the efficiency of the proposed protocols, the performance is evaluated against various parameters such as number of nodes, speed and pause time of nodes. The ease of implementation of routing protocols is also compared. The comparison shows that all the protocols perform similarly before ε-value correlation is used on σ-bandwidth routes. The network instantly recognizes the attack and focuses on attack-free routing. If the attack persists, all the proactive attack-free routing protocols find the attacked nodes and preemptively isolate them. So, a sophisticated attack on σ-bandwidth can go undetected by the σ-Max protocols and can take control of the entire route discovery. However, proposed hybrid protocols perform well since they can keep the network alive with attack-free β-bandwidth routes. The existence of so many benign β-bandwidth routes and the difference in attack cost on σ-Max and β-Max routes helps the proposed hybrid protocols to do better than any other protocols.

## 9. Discussion

The discussion section summarizes the implications of the research findings for the security of MANETs, discussing the main considerations, and outlines several possible avenues for future research in each domain.

A number of considerations should accompany the selection and design of any hybrid secure routing protocol employing the techniques proposed in this research. First, the Mobile Ad-hoc Network (MANET) domain for which the protocol is developed should be clearly delimitated, with particular attention given to the nature of the devices, the mobility model, the network size and density, the routing metrics that may be used, and so forth. The impact and viability of proposed techniques may differ markedly depending upon such considerations. Second, the sensitivity of a protocol's performance to alterations of a few key parameters should be disclosed, particularly when differences in performance could change the potential applicability of this protocol. For example, protocols may need to be tuned to best suite MANETs in a very specific set of conditions. Of course, it is also important that the scaling characteristics of the proposed protocols be discussed. Third proposed hybrid protocols can be attacked in a variety of ways. Therefore possible future improvements and countermeasures should be stated clearly, allowing users to make informed decisions about the applicability of a protocol in particular contexts [9].

In addition to the above considerations, research on the hybrid secure routing of MANETs more generally should also continue along the following avenues: (i) Proposals for hybrid secure routing protocols should be developed that employ the techniques that were found to be effective during this research and that consider security issues related to them. In particular, the function, compatibility, and efficacy of the trust management, cryptographic key, false information, and incentive techniques should be investigated in combination. (ii) Further analyses of existing hybrid secure routing protocols should be undertaken. In particular, the function, compatibility, and efficacy of the hybrid secure routing techniques in trace (simulation) analyses of the MHDH protocols should be explored. (iii) Wider security issues related to hybrid secure routing protocols should be investigated. In particular, routing protocols should be protected from or rendered unaffected by certain secondary forms of attacks (e.g., attacks on authentication processes). In so doing, these hybrid protocols will be made more complementary [1].

### 9.1. Implications for MANETs Security

The implications of this research for the improved security of MANETs are discussed in this section. Research issues relating to the possible impact of findings, predicted developments, potential applications and practical implications are addressed. In recent years, there has been growing interest in establishing secure MANET routing protocols. Towards this end, two hybrid secure routing protocols GSHARP and GSHARPP within the context of the Ad-Hoc On-Demand Distance Vector routing protocol have been proposed. As part of this research, security enhancements have been proposed to overcome spoofing, sinkhole attacks and route replay attacks [1].

It is intended these protocols be used as part of infrastructure less, highly mobile wireless networks in real-world applications such as emergency response to disaster situations. The environments are characterized by the need to quickly organize and share information/communication among first responders while being open to eavesdropping and deliberately posed network attacks. The purpose of the research is to improve the security level of MANETs and make it more challenging for a malicious agent to disrupt or seize control of a MANET routing protocol. Furthermore, it is proposed to make these protocols practical for use in real-world situations by keeping the additional cryptographic overhead low. The research was conducted using a standard simulation environment where communications are simulated as being transmitted over wireless channels in addition to differences in protocol performance under various context conditions [9].

### 9.2. Future Research Directions

Several avenues for future research and development are proposed in the quest for hybrid secure routing protocols for MANETs.

Due to constantly changing networks in MANETs, topology changes such as breakage and formation of new links happen at a high rate. As a result, it is difficult for traditional routing protocols to sustain a fixed and working route for a certain interval of time that is acceptable for nodes to communicate with each other. There is a need for hybrid routing protocols in which the nodes of the network can switch between proactive and reactive protocols based on current requirements. In this way, problem of route stability is taken care of.

Security problems associated with MANETs are enormous due to the shared wireless channel. Certain vulnerabilities are present in the protocols that can be exploited by an attacker. There is a need to investigate ECDSA (Elliptic Curve Digital Signature Algorithm) based secure routing protocols for MANETs different from the trust-based hybrid routing protocol proposed. Handling flooding, black hole and rushing attacks effectively and efficiently is a difficult task and requires novel techniques [9].

## 10. Conclusion and Recommendations

The escalating security threats in Mobile Ad-Hoc Networks (MANETs) due to their open, infrastructure-less and dynamic nature have placed substantial importance upon the role of secure routing protocols. The simple hope that security measures will work implicitly and invisibly is no longer useful. It has become clear that security has to be engineered at every layer, beginning from the bottom-most layer, as each layer is tied to its neighboring layer(s) in terms of functionality, addressing schemes, metrics, and topologies. In addition, each layer is constantly threatened from the top-most layer down to the Data Link layer [1].

In response to the critical requirement of security in communications, a wide variety of protocols often referred to as security protocols have been developed. These security protocols carry on one or more security functions. Security protocols may be divided broadly into three categories, namely link-level security protocols which offer security features between directly-linked network neighbors, end-to-end security protocols which provide security features between the network endpoints, and hybrid security protocols that take advantage of both link-level and end-to-end security protocols. This study reviews the existing hybrid secure routing protocols to understand their approaches and to give recommendations for implementation [8].

### 10.1. Summary of Key Findings

A Mobile Ad-Hoc Network (MANET) is a collection of mobile nodes and devices that can communicate over wireless links without any fixed infrastructure. Depending on the types of routing protocols, there are three types of routing protocols for MANETs like: proactive, reactive, and hybrid. A recent trend in Ad Hoc network routing is the reactive on-demand philosophy where routes are established only when required. This paper presents Hybrid Secure Routing Protocol (HSRP) for MANET addressing both security and performance issues. Namely, it is a combination of on-demand reactive and distance vector proactive AOMDV routing protocol. In reactive phase, node discovery and route establishment is performed in AODV fashion but as soon as routing table is created, it periodically maintains the routes via proactive updating using AOMDV routing protocol. In both phases, nodes participate in a multi-signature-based authentication mechanism to provide security against malicious nodes like blackhole, flooding, and sinkhole [1]. An efficient method and mathematical models are proposed to evaluate the performance impact of network traffic, hop count, and time delay on the routing performance in both phases of HSRP. This performance evaluation is done in MANET and is best suited for hybrid secure routing in mobile ad-hoc networks. Several experiments have been conducted on benchmark scenarios involving up to 100 nodes using Network Simulator (NS 2) to corroborate the correctness of proposed scheme and confidence in analytical model.

Numerous attacks on routing protocol in Ad-Hoc network affect particularly intermediate nodes which are responsible for packet forwarding. Security and protocols in MANETS have been recently focused due to the challenges posed by their unique characteristics especially wireless interface, mobile topology and lack of infrastructure [9]. These challenges need to be addressed and modeled in order to design secure, reliable and energy efficient protocols for mobile ad-hoc networks. Security and reliable protocols in MANETS need to be properly analyzed, scrutinized, modeled and fully investigated before becoming standard in the mobile industry. A model to evaluate the performance impact of traffic type or application, network size, mobility, hop count, time delay, topological change and skewed traffic on the performance of routing protocols in mobile ad-hoc networks have been proposed and is also applied to routing protocols.

### 10.2. Recommendations for Implementation

Considerable emphasis has been placed on the implementation and deployment of hybrid secure routing protocols in the domain of Mobile Ad-Hoc Networks (MANETs). These recommendations aim to provide actionable insights and guidance for the practical adoption and integration of the outcomes derived from the research conducted.

Initially, it is prudent to ensure that all mobile nodes participating in the MANET have the requisite hardware and software specifications to support the routing protocols [1]. Within the mobile nodes, there are necessary specifications for the software implementation, which consists of the code for both the Ad-hoc On-Demand Distance Vector (AODV) routing protocol and the hybrid secure routing protocol. All nodes must have the same versions of these codes, as incorrect operation will be more likely if nodes with differing versions execute the same protocol. The absence of either code in a mobile node will render that node incapable of participating in that routing protocol [13].

Moreover, the hardware requirements include having at least a Network Interface Card (NIC) and radio transmission range capable of supporting wireless packet communication to and from a shared channel. For mobile nodes to communicate with ones outside their transmission range, all nodes need to remain within the transmission ranges of each other. Transmission ranges should also overlap to ensure complete connectivity. Each node should have a unique network address such as an Internet Protocol (IP) address, which will not permit nodes to communicate freely with one another.

[1]A Mobile Adhoc Network (MANET) is characterized by mobile nodes, multihop wireless connectivity, infrastructure less environment and dynamic topology. A recent trend in Ad Hoc network routing is the reactive on-demand philosophy where routes are established only when required. Stable Routing, Security and Power efficiency are the major concerns in this field. This paper is an effort to study security problems associated with MANETS and solutions to achieve more reliable routing. The ad hoc environment is accessible to both legitimate network users and malicious attackers. The study will help in making protocol more robust against attacks to achieve stable routing in routing protocols.

[2]Similar to denial of service attacks on internet application servers such as HTTP and FTP. MANET can be corrupted by overwhelming radio frequency (RF) signal. Distributed DoS attack is a more severe threat: if the attackers have enough computing power and bandwidth, smaller MANETs can be crashed or congested very easily. Radio jamming and battery exhaustion are two ways in which nodes cannot communicate with each other. If authentication of nodes is not supported, malicious nodes can be able to join the network without detection, send false routing information, and masquerade as some other trusted node. There are three kinds of fabrication attacks. To generate route error messages. To corrupt routing information. Other fabrication attacks. In any case, these kinds of attacks are not easy to detect. In data flooding, the attacker will send unwanted data items to congest the network. The attacker will send a large amount of RREQ requests to waste the bandwidth and resources of the network, usually the destination IP chosen for RREQ will not exist in the network. Any destination node will always be busy in receipt of unwanted data. This paper provides an overview of the security issues in MANETs. It classifies the attacks that are possible against the existing routing protocols. An understanding of these attacks and their impacts on the routing mechanism will help researchers in designing secure routing protocols.

## 11. References

References:

[1] N. (Nidhi) Goyal and S. (Susheel) Kumar, "Proposed Scheme for Secured Routing in MANET," 2017. [PDF]

[2] P. Kakkar and K. Kumar Saluja, "Vulnerabilities for Reactive Routing in Mobile Adhoc Networks," 2015. [PDF]

[3] J. Sen, "A Multi-Path Certification Protocol for Mobile Ad Hoc Networks," 2012. [PDF]

[4] A. Rajaram and D. S. Palaniswami, "A Trust Based Cross Layer Security Protocol for Mobile Ad hoc Networks," 2009. [PDF]

[5] V. Toubiana and H. Labiod, "ASMA: towards adaptive secured multipath in MANETs," 2012. [PDF]

[6] A. R.M. Shabut, K. P. Dahal, S. K. Bista, and I. U. Awan, "Recommendation based trust model with an effective defence scheme for MANETs," 2015. [PDF]

[7] M. Riyaz Belgaum, S. Musa, M. Mohd Su'ud, M. Alam et al., "Secured Approach Towards Reactive Routing Protocols Using Triple Factor in Mobile Adhoc Networks," 2019. [PDF]

[8] E. A. Panaousis, G. Drew, G. P. Millar, T. A. Ramrekha et al., "A Testbed Implementation for Securing OLSR in Mobile Ad hoc Networks," 2010. [PDF]

[9] S. Gour and S. Sharma, "A Survey of Security Challenges and Issues in Manet," 2015. [PDF]

[10] A. Vijaya Kumar and A. Jeyapal, "Self-Adaptive Trust Based ABR Protocol for MANETs Using Q-Learning," 2014. ncbi.nlm.nih.gov

[11] J. Sen, "A Distributed Trust Management Framework for Detecting Malicious Packet Dropping Nodes in a Mobile Ad Hoc Network," 2010. [PDF]

[12] S. Krishna Chimbli Venkata, "Sector Based Clustering & Routing," 2004. [PDF]

[13] K. Majumder and S. Kumar Sarkar, "Hybrid Scenario Based Performance Analysis of DSDV and DSR," 2010. [PDF]

[14] A. Aggarwal, S. Gandhi, and N. Chaubey, "Performance Analysis of AODV, DSDV and DSR in MANETs," 2014. [PDF]

[15] P. Ghee Lye and J. C. McEachen, "A Comparison of Optimized Link State Routing with Traditional Ad-hoc Routing Protocols," 2006. [PDF]

[16] Y. Cheng, "Performance Analysis of Transactional Traffic in Mobile Ad-hoc Networks," 2014. [PDF]

# Improving Recommendation quality via Ensemble Neural Networks

Ramzi Khantouchi[1,*], Ibtissem Gasmi[1], Djaber Abbas[1] and Abderaouf Bahi[1]

[1]*Computer Science and Applied Mathematics Laboratory, Chadli Bendjedid University, El Tarf 36000, Algeria*

**Abstract**

Recommender systems are essential in e-commerce platforms for predicting user preferences and delivering personalized recommendations. However, addressing the sparsity of user-item interactions and capturing dynamic user behavior remains a challenge. This study proposes an ensemble neural network approach combining Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Multilayer Perceptron (MLP) to improve recommendation quality and tackle data sparsity. By leveraging LSTM's ability to capture temporal dependencies, GRU's computational efficiency, and MLP's non-sequential feature learning, the ensemble mitigates individual model limitations. Experiments on the MovieLens and Amazon Beauty datasets show that the ensemble outperforms standalone models, achieving up to 2.8% higher Precision and Recall and 2.8% lower RMSE. These results demonstrate the ensemble's capability to provide robust and accurate recommendations, even in sparse data environments, enhancing user satisfaction.

## 1. Introduction

Recommender systems are extensively utilized in e-commerce platforms such as Amazon, Alibaba, eBay, and Taobao to predict users' potential interests from a large pool of items. In practical scenarios, user behaviors are dynamic and constantly evolving, necessitating the accurate characterization of these behaviors to provide effective recommendations—a fundamental objective of recommender systems. While these systems significantly facilitate the dissemination and accessibility of information, they also raise serious privacy concerns due to the often intrusive collection of demographic and behavioral data from users [1]. The unintentional exposure of sensitive information poses ethical issues and risks violating data protection regulations, such as the European Union's General Data Protection Regulation (GDPR) [2]. However, reducing the volume of user data collected to address privacy concerns can severely affect recommendation quality, leading to the well-known issue of data sparsity. To address data sparsity, recent research has explored both enhancements to traditional collaborative filtering (CF) methods and the integration of deep learning techniques [3, 4, 5]. For example, the work in [3] proposed an item-based CF algorithm that utilizes Kullback–Leibler (KL) divergence to

*2nd International Conference on Autonomous Systems and their Applications (ICASA'24) , December 15, 2024, El Tarf, Algeria*

[*] Corresponding author.

✉ r.khantouchi@univ-eltarf.dz (R. Khantouchi); gasmi-ibtissem@univ-eltarf.dz (I. Gasmi); dj.abbas@univ-eltarf.dz (D. Abbas); a.bahi@univ-eltarf.dz (A. Bahi)
🌐 https://ramzikhantouchi.github.io/ (R. Khantouchi)
iD 0009-0009-5391-9022 (R. Khantouchi); 0000-0002-8939-1727 (I. Gasmi)

refine item similarity measurements. This approach improved similarity accuracy and integrated more rating information into prediction processes, enhancing recommendation quality in sparse datasets. Similarly, [6] focused on mitigating sparsity and cold-start challenges in CF by refining similarity measures such as Cosine similarity, Pearson correlation, and Adjusted Cosine similarity, resulting in improved recommendation accuracy. Other studies have proposed hybrid and attribute-based approaches. For instance, [7] introduced a hybrid CF model for consumer service recommendations in mobile cloud environments, leveraging user preferences to mitigate sparsity issues and improve accuracy. Meanwhile, [5] proposed the Attribute-based Neural Collaborative Filtering (ANCF) model, which incorporates user and item attributes into CF using an attention mechanism and a multi-layer perceptron. This model effectively captures the varying impacts of attributes, providing a comprehensive feature representation that enhances recommendation performance, as validated on multiple datasets.

Further advancements include leveraging genre information and contextual data to enhance recommendations. For example, [8] developed an item-based CF algorithm that integrates genre information and accounts for dynamic changes in user preferences, while [9] combined CF with genetic algorithms to utilize contextual user and item data. Additionally, [10] proposed a neural recommendation model designed for non-independent and identically distributed (Non-IID) data, which integrates explicit and implicit interactions within CF frameworks. Authors in [11] addressed the sparsity problem by leveraging the power of Large Language Models (LLMs). They used GPT-4o as a recommendation model to enhance recommendation quality. They formatted the previous user interaction sequence as text and provided the candidate item as input. The recommendation model then outputs the probability of the candidate item being relevant to the user. These studies collectively demonstrate the potential of diverse methodological advancements in addressing the challenges of data sparsity and enhancing recommendation accuracy. Authors in [12] proposed a hybrid approach that integrates neighborhood-based techniques with neural network model-based methods. The outputs of these models are fused using an additional neural network, creating a unique framework. Authors in [13] utilized group classification and ensemble learning techniques to enhance prediction accuracy in recommender systems. A key challenge addressed was user analysis, which becomes complex and resource-intensive due to the large-scale data and user base. To tackle this, graph embedding was proposed as a solution, enabling the simulation of user behavior through vector generation. This approach effectively simplified user behavior analysis while maintaining high efficiency. The study classified users similar to the target user using ensemble learning, fuzzy rules, and decision trees, subsequently providing personalized recommendations through a heterogeneous knowledge graph and embedding vectors. Evaluated on the MovieLens datasets, the proposed method demonstrated superior efficiency and performance.

In this paper, we propose a novel ensemble neural network framework designed to enhance recommendation systems, in sparse scenarios. Our approach combines three powerful neural network architectures—Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Multilayer Perceptron (MLP)—into a single ensemble model to leverage their complementary strengths. By combining these models, the ensemble approach benefits from each model's unique capabilities, improving the overall performance of the recommendation system. This method not only increases the accuracy of predictions but also addresses the critical problem of data sparsity, enabling more effective personalization.

The main contributions of this study include the following:

- We introduce a new ensemble model that combines LSTM, GRU, and MLP, each contributing to different aspects of the recommendation process. This hybrid approach enhances predictive accuracy, particularly in sparse datasets
- Evaluate the performance of the proposed approach on two benchmarks datasets.
- Use multiple evaluation metrics such as precision, recall, F1 score, MAE, RMSE and MSE to asses the performance of the proposed approach.

## 2. Methodology

### 2.1. Long Short-Term Memory

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture specifically designed to overcome the limitations of traditional RNNs, particularly the vanishing gradient problem. LSTM introduces memory cells that can maintain information over long time steps, making it effective at learning long-range dependencies in sequential data. The LSTM unit is controlled by three gates: the input gate, which decides which new information to store; the forget gate, which determines what information to discard; and the output gate, which controls how much of the stored memory should be used to generate the output. This gating mechanism enables LSTMs to capture patterns in time series, natural language, and other sequential tasks where retaining information over extended periods is crucial.

### 2.2. Gated Recurrent Unit

Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) architecture designed to capture dependencies in sequential data, similar to Long Short-Term Memory (LSTM), but with a simpler structure. GRUs use two gates: the update gate, which determines how much of the previous information should be retained and how much of the new information should be incorporated, and the reset gate, which controls the amount of past information to forget. Unlike LSTM, GRU combines the hidden state and the memory cell into a single component, reducing the number of parameters and making it computationally more efficient. It allows GRUs to achieve comparable performance to LSTMs while often converging faster, making them a popular choice for various sequence modeling tasks such as natural language processing and time series prediction.

### 2.3. Multilayer Perceptron

Multilayer Perceptron (MLP) is a type of feedforward neural network composed of multiple layers of neurons, including an input layer, one or more hidden layers, and an output layer. Each neuron in an MLP processes information by computing a weighted sum of its inputs and applying a nonlinear activation function. This structure allows the MLP to learn complex, non-linear relationships in the data. MLPs are fully connected, meaning each neuron in one layer is connected to every neuron in the next layer. They are trained using the backpropagation algorithm, adjusting the weights to minimize the error between predicted and actual outcomes. MLPs are effective for tasks like classification and regression in domains such as image recognition and tabular data analysis, where the input data is not sequential and does not have temporal dependencies.

### 2.4. Ensemble Neural Networks

Ensemble Neural Networks refer to a technique where multiple neural network models are combined to improve overall performance, stability, and generalization capabilities compared to a single model. The basic idea behind ensemble methods is to leverage the diversity of predictions from different models to reduce variance, prevent overfitting, and achieve more robust predictions. In ensemble neural networks, different architectures or the same architecture trained on different subsets of data are used, and their predictions are aggregated through techniques like averaging, voting, or stacking. Common types of ensemble methods include bagging, where models are trained independently on different random samples of the data, and boosting, where models are trained sequentially to correct errors made by previous models. Ensemble methods are widely used in applications such as classification, regression, and even time series prediction, where combining models often leads to more accurate and reliable results compared to using a single neural network.

### 2.5. Approach

This study aims to explore the effectiveness of using Ensemble Neural Networks, specifically combining Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), and Gated Recurrent Unit (GRU) architectures, for item recommendation in e-commerce platforms. By leveraging the unique strengths of each model, this ensemble approach seeks to enhance the accuracy and relevance of recommendations presented to users. LSTMs excel in capturing sequential dependencies in user behavior and interactions over time, making them suitable for understanding temporal patterns in user preferences. In contrast, MLPs can effectively learn complex relationships from user and item features, while GRUs offer a more computationally efficient option for modeling sequential data. The integration of these models through techniques such as stacking or averaging will enable the ensemble system to mitigate the weaknesses of individual models and improve generalization across diverse user profiles. The research will investigate various configurations of the ensemble, evaluate their performance against traditional recommendation systems, and analyze the potential benefits of this hybrid approach in providing personalized, context-aware recommendations that can enhance user satisfaction and engagement.

## 3. Experiments

### 3.1. Datasets

We select two popular datasets to evaluate our performance on both small and large scales, with their statistics presented in Table (1). (1) **MovieLens**: MovieLens is a platform that recommends movies to users based on their past ratings and is now one of the most frequently utilized benchmarks in the field of recommender systems. In our experiments, we utilize MovieLens-100k. (2) **Amazon Beauty**: The online reviews and ratings from Amazon are widely recognized benchmarks for assessing recommendation algorithms. This dataset encompasses a wide range of beauty and personal care products, including skincare, cosmetics, and hair care items, and includes detailed user-generated content such as reviews, star ratings, and feedback on product efficacy.

**Table 1**
Statistics of the datasets

| Data | Movielens | Beauty |
|---|---|---|
| # users | 943 | 1,210,271 |
| # items | 1,682 | 249,274 |
| # interactions | 100,000 | 2,023,070 |

## 3.2. Evaluation Metrics

Standard metrics, such as Recall (Equation (1)), Precision (Equation (2)), and F1 Score (Equation (3)), are commonly used to evaluate the performance of recommendation systems by measuring the relevance and accuracy of the recommended items. In addition to these ranking-based metrics, we also employ error metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to assess the accuracy of the model's predicted relevance scores. MSE (Equation (4)) captures the average squared difference between the predicted and actual relevance values, penalizing larger errors more heavily. RMSE (Equation (5)) offers a more interpretable version by taking the square root of MSE, which helps in comparing models. MAE (Equation (6)), on the other hand, calculates the average absolute difference between predicted and actual relevance scores, providing a more direct measure of the prediction error. These metrics enable a comprehensive evaluation of the recommendation model's ability to rank items effectively.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{6}$$

TP, TN, FP, FN, $y_i$, $\hat{y}_i$ and N represent True Positive, True Negative, False Positive, and False Negative, the actual relevance score, the predicted relevance score and the total number of observations respectively.

### 3.3. Experimental Setup

To evaluate the proposed ensemble approach, we compare it with individual models: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Multilayer Perceptron (MLP). The models are trained on the **MovieLens-100k** and **Amazon Beauty** datasets, described in Table (2). For the ensemble, we employ a stacking method, where predictions from LSTM, GRU, and MLP are aggregated using a meta-learner trained on the validation set predictions.

Each model was implemented using PyTorch and optimized using the Adam optimizer with a learning rate of $10^{-3}$. We employed early stopping with a patience of 5 epochs to avoid overfitting. The data was split into training (80%), validation (10%), and test (10%) sets, ensuring no overlap in users between splits.

### 3.4. Performance Results

The performance of the models is evaluated using ranking metrics (Precision, Recall, and F1 Score) and error metrics (MSE, RMSE, and MAE). The results are summarized in Table (2).

**Table 2**
Performance Comparison of Models on the Test Set

| Dataset | Metric | LSTM | GRU | MLP | Ours |
|---|---|---|---|---|---|
| MovieLens | Precision | 0.745 | 0.728 | 0.710 | **0.768** |
| | Recall | 0.652 | 0.634 | 0.622 | **0.671** |
| | F1 Score | 0.695 | 0.678 | 0.663 | **0.716** |
| | RMSE | 0.831 | 0.849 | 0.860 | **0.810** |
| | MAE | 0.671 | 0.682 | 0.694 | **0.645** |
| Beauty | Precision | 0.683 | 0.670 | 0.650 | **0.702** |
| | Recall | 0.590 | 0.578 | 0.562 | **0.612** |
| | F1 Score | 0.633 | 0.621 | 0.602 | **0.653** |
| | RMSE | 0.927 | 0.940 | 0.958 | **0.901** |
| | MAE | 0.711 | 0.723 | 0.734 | **0.698** |

### 3.5. Discussion

The experimental results demonstrate that the ensemble model outperforms the individual models (LSTM, GRU, and MLP) across both datasets, achieving the highest Precision, Recall, and F1 Score, which underscores its effectiveness in ranking relevant items. Specifically, the ensemble model improves Precision by 2.3% and 2.8% on the MovieLens and Beauty datasets, respectively, compared to the best-performing individual model (LSTM), and similar gains are observed in Recall, highlighting its ability to recommend a larger proportion of relevant items without compromising precision. Furthermore, the ensemble achieves lower RMSE and MAE values, with its RMSE on the Beauty dataset (0.901) being 2.8% lower than that of LSTM (0.927), indicating that its predictions align more closely with actual ratings. These improvements stem from the ensemble's ability to leverage the strengths of LSTM (capturing temporal dependencies), GRU (computational efficiency), and MLP (non-sequential feature learning), effectively mitigating the individual weaknesses of these models. For instance, while MLP struggles with sequential dependencies, its strength in learning feature interactions complements the sequential modeling capabilities of LSTM and GRU, resulting in a more robust and accurate recommendation system.

## 4. Conclusion

This paper introduced an ensemble neural network framework to address the sparsity problem and improve the accuracy of recommendations in e-commerce platforms. By combining the strengths of LSTM, GRU, and MLP, the proposed approach effectively tackled the challenges of sparse user-item interactions and dynamic user behavior. Experimental results on the MovieLens and Amazon Beauty datasets revealed that the ensemble consistently outperformed individual models, achieving significant gains in ranking metrics (Precision, Recall, F1 Score) and reductions in error metrics (RMSE, MAE). The ensemble's ability to handle sparse data and capture diverse patterns in user behavior underscores its potential for broader applications in recommender systems. Future research will focus on further optimizing the ensemble model and evaluating its effectiveness in other domains with sparse and complex data structures.

## References

[1] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, Y. Zhang, Membership inference attacks against recommender systems, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 864–879.

[2] C. J. Hoofnagle, B. Van Der Sloot, F. Z. Borgesius, The european union general data protection regulation: what it is and what it means, Information & Communications Technology Law 28 (2019) 65–98.

[3] W. Zhao, H. Tian, Y. Wu, Z. Cui, T. Feng, A new item-based collaborative filtering algorithm to improve the accuracy of prediction in sparse data, International Journal of Computational Intelligence Systems 15 (2022) 15.

[4] A. Althbiti, R. Alshamrani, T. Alghamdi, S. Lee, X. Ma, Addressing data sparsity in collabo-

rative filtering based recommender systems using clustering and artificial neural network, in: 2021 IEEE 11th annual computing and communication workshop and conference (CCWC), IEEE, 2021, pp. 0218–0227.

[5] H. Chen, F. Qian, J. Chen, S. Zhao, Y. Zhang, Attribute-based neural collaborative filtering, Expert Systems with Applications 185 (2021) 115539.

[6] C. Ajaegbu, An optimized item-based collaborative filtering algorithm, Journal of ambient intelligence and humanized computing 12 (2021) 10629–10636.

[7] Q. Zhou, W. Zhuang, H. Ren, Y. Chen, B. Yu, J. Lou, Y. Wang, Hybrid collaborative filtering model for consumer dynamic service recommendation based on mobile cloud information system, Information Processing & Management 59 (2022) 102871.

[8] I. Gasmi, H. Seridi-Bouchelaghem, L. Hocine, B. Abdelkarim, Collaborative filtering recommendation based on dynamic changes of user interest, Intelligent Decision Technologies 9 (2015) 271–281.

[9] I. Gasmi, F. Anguel, H. Seridi-Bouchelaghem, N. Azizi, Context-aware based evolutionary collaborative filtering algorithm, in: International Symposium on Modelling and Implementation of Complex Systems, Springer, 2020, pp. 217–232.

[10] M. F. Aljunid, M. D. Huchaiah, Integratecf: Integrating explicit and implicit feedback based on deep learning collaborative filtering algorithm, Expert Systems with Applications 207 (2022) 117933.

[11] R. Khantouchi, I. Gasmi, A. Bahi, Enhancing recommendation quality in sparse scenarios using large language models (llms) (2024).

[12] P. Sahu, S. Raghavan, K. Chandrasekaran, Ensemble deep neural network based quality of service prediction for cloud service recommendation, Neurocomputing 465 (2021) 476–489.

[13] S. Forouzandeh, K. Berahmand, M. Rostami, Presentation of a recommender system with ensemble learning and graph embedding: a case on movielens, Multimedia Tools and Applications 80 (2021) 7805–7832.

# Improving Scalability in Blockchain based Federated Learning Using Off-chain Storage

Ala Djeddai[1,*,†], Rofaida Khemaissia[2,†] and Makhlouf Derdour[3,†]

[1]*Laboratory of Computer Science and Applied Mathematics (LCSAM), Chadli Bendjedid El-Tarf University, B.P 73, El Tarf 36000, Algeria*

[2]*Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa,12002, Algeria*

[3]*LIAOA Laboratory, Larbi Ben M'Hidi University, Oum El Bouaghi, Algeria*

### Abstract

Owing to the ubiquity of user data everywhere, preserving privacy has remained a problem for decades, where AI requires raw data to accomplish machine learning models. Federated learning was first launched by Google to build machine learning models based on datasets distributed across multiple devices while preventing data leakage. We introduce MFchain, a new blockchain-based federated learning design that differs from most approaches by using off-chain storage to manage the global model and preserve its privacy. Access control was ensured to prevent illegal model modifications. We use Hyperledger Fabric, a permissioned blockchain that demonstrates the efficiency and feasibility of our framework by implementing smart contracts.

### Keywords

Blockchain, Hyperledger Fabric, Data Integrity, Smart Contracts, Privacy, Certificate attestation, and verification.

## 1. Introduction

Old machine learning requires centralizing training data; however, in 2016, Google launched a new concept that allows Android mobile phone users to update models locally without breaching their private personal data [6]. Federated learning (FL) builds machine learning models based on datasets that are distributed across multiple devices while preventing data leakage [9]; in May 2018, the GDPR was enforced by the European Union to protect users' personal data. FL has given the opportunity for users to train their own data without the need for a third party, whereas, after the end of all rounds, the central server uses an algorithm for aggregating the local model to obtain a global model and shares another time with the new global model for starting another round, the interaction between the central server and the participants (users) has to be in a secure manner; thus, multiple approaches have used different cryptographic methods to ensure both security and privacy of machine learning models, for example, secret multi-party computation, differential privacy, homomorphic cryptographic encryption, and so on. In the same endeavor, Blockchain, the famous technology that has been integrated to maintain both security and privacy in FL; it is counted as a secure central server that assures the FL without participants' or model privacy leakage; many studies put forward the concern of preserving privacy using blockchain with another cryptographic method; the main issue is scalability (block size) when using blockchain as a decentralized server; at the same

[*] Corresponding author.

[†] These authors contributed equally.

a.djeddai@univ-eltarf.dz (A. Djeddai); khemaissia.rofaida@univ-tebessa.dz (R. Khemaissia); derdour.makhlouf@univ-oeb.dz (M. Derdour)

0000-0002-9354-4108 (A. Djeddai); 0000-0002-4092-537X (R. Khemaissia); 0000-0001-6622-4355 (M. Derdour)

time, a distributed database we handle with this problem by proposing an MFchain framework that handles scalability problems by using off-chain storage to maintain the model's privacy.

In this paper, we propose MFchain, a new design that treats the problems discussed above by using a permissioned blockchain hyperledger fabric that cooperates with off-chain storage. In our design, blockchain is used for authentication by applying public-key cryptography. The main contributions of this study are as follows:

- Our work proposes a new federated learning privacy preserving method using blockchain for protecting the privacy of both the model and participants.
- This work utilizes a permissioned blockchain as a mechanism for authentication, access control, and logging using smart contracts. We configured an orchestrator for managing the system by validating the identities and updating the blockchain.
- Offchain storage is used to improve the scalability of blockchain-based federated learning. Therefore, only important data regarding the global model are stored in the BC.

This paper is structured as follows: Section 2 presents the background and related work. Section 3 describes the proposed MFChain approach. Section 4 presents the implementation of the MFchain. The evaluation results and security analysis are presented in Section 5. Finally, Section 6 concludes the study.

## 2. Background and Related works

This section describes some essential terminologies that will be used in this study in addition to the related approaches that contribute to building our idea.

### 2.1. Hyperledger Fabric (HLF)

Fabric is an open-source permissioned blockchain system, it is an enterprise-grade open-source blockchain platform hosted by the Linux Foundation [1], HLF is a modular, decentralized ledger technology (DLT) platform that was designed by IBM for industrial enterprise use. The key features of Hyperledger fabric are an open source, permissioned, governance, and access control, and its performance, the key components of HLF as follows:

#### 1) Membership Service Provider (MSP)

is used to manage identities on the blockchain network and is used to authenticate clients who want to join the blockchain network. It plays the function of middleman between organizations that want to participate in a network and those that want to use their services.

#### 2) Nodes

HLF has three different nodes; each node has a particular role and functionality, and they could be a client, peer, or an order. A client in HLF is used to issue a transaction in the network; nevertheless, a peer manages ledgers and smart contracts. Peers can be endorsers or committers. The endorser is in charge of executing the chaincode to endorse transaction proposals, and a committer updates the ledger state by validating and committing transactions. The third node is Orderer, also known as an "ordering node, " which performs this transaction ordering, which, along with other Orderer nodes, forms an ordering service.

#### 3) Chaincode:

This is a program written in Go, Node.js, or Java. It is referred to as a smart contract, which initializes and manages the ledger state through transactions submitted by applications.

#### 4) Channel

It is a private subnetwork between two or more specific network members, for the purpose of conducting private and confidential transactions.

### 2.2. Related Works

Fedchain is referred to as a set of contributions and approaches that attempt to protect federated learning based on blockchain, which aims to increase user and data privacy. El Rifai et al. [3] first integrated federated learning for machine learning models and blockchain technologies in the medical field; they proposed a Smart Contract (SC) implementation of a coordinating server for an FL algorithm. Similarly, Abou Elhouda et al. [4] designed a HealthFed blockchain and federated learning-based platform to preserve privacy and distribute the learning process between different clinic collaborators, and HealthFed ensures a secure aggregation of local model updates by leveraging a secure multiparty computation scheme. In the same endeavor Sun et al. [5] built a robust secure federated learning platform based on a permissioned Blockchain Hyperledger Fabric and proposed that BC played the role of a central server where individual local updates (local training parameters) are encrypted based on threshold homomorphic encryption and then recorded on a distributed ledger. Undoubtedly, the aforementioned approaches are efficient and feasible solutions for preserving user privacy and security simultaneously. However, some of them considered the blockchain as a distributed server and benefited from the decentralized feature to store the updated model after the aggregation process, which, in reality, could turn into a drawback. Owing to the size of the data stored (block size), this paper presents a healthcare federated learning based on blockchain technology to preserve user privacy that does not need to store updated data on blockchain to fetch a platform that is more scalable.

## 3. Proposed method

### 3.1. System Design

### 3.1.1. Identity Management

The membership service provider is considered as trust authority in Blockchain which is responsible for generating participants' identities (healthcare institutes for instance), the key pair is generated as a tuple $(pk_i, sk_i)$ ,we can consider the public key $pk_{\ i}$as a pseudonym for validating the authentication process. We have generated the $(pk_i, sk_i)$ by using the RSA algorithm or ECC, where the $pk_{\ i}$of an entity is saved in a dataset. However, the private key $sk_i$ remains as a secret for the participant.

### 3.1.2. MFchain Architecture

The proposed federated learning architecture is described in detail in this section. MFChain is a decentralized and distributed platform for federated learning based on the blockchain technology. This is different from most traditional federated-chain frameworks, which increase the security and privacy of the data, as well as owner privacy. The main objective of MFchain is to securely carry out any type of learning Machine or Deep learning…) and store the en-pointer of the results on the blockchain. The Blockchain plays the role of the central server; instead of using a single point of failure, it can behave as a decentralized server to avoid any type of dominating attack (Sybil attack, Byzantian fault-tolerant attack, etc.). Nevertheless, we use secure off-chain storage, for instance (MongoDB) on purpose of store the model parameters and keep only the en-pointer on the on-chain, which enables only legitimate participants to query the updated model.

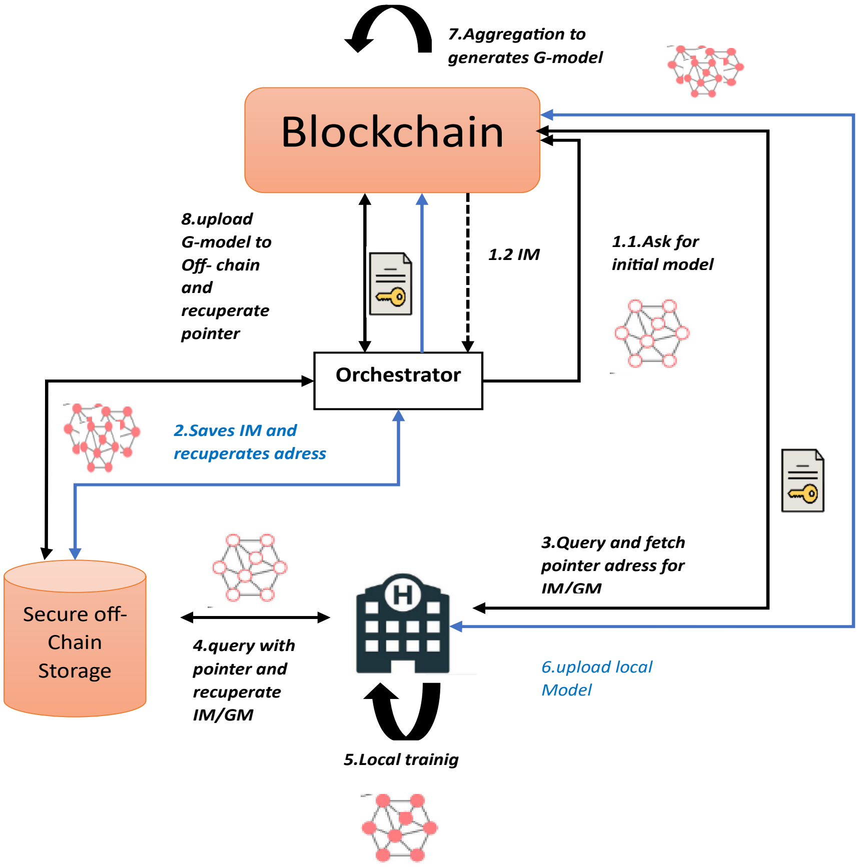

7.Aggregation to generates G-model
Blockchain
8.upload G-model to Off- chain and recuperate pointer
1.2 IM
1.1.Ask for initial model
Orchestrator
2.Saves IM and recuperates adress
3.Query and fetch pointer adress for IM/GM
Secure off-Chain Storage
4.query with pointer and recuperate IM/GM
6.upload local Model
5.Local trainig

**Figure 1:** A complete training process of one round of our MFchain

MFchain has four main components that interact with them in distributed and decentralized networks. A group of participants permissioned blockchain, for instance HLF, a secure off-chain storage, and orchestrator.

### A) Participants:

Workers are external entities; for instance, healthcare institutes (hospitals, research clinics, etc.) are in charge of gathering the raw data in order to fulfill the local training and upload the updated parameters.

### B) Permissioned Blockchain

For instance, Hyperledger Fabric which plays the role of central server, it is in charge of access management, generating the initial model and accomplishing the global model (G-model). In addition, it is responsible for preserving model privacy.

### C) Off-chain storage:

is utilized to save the model's initial parameters (IM) or global model (GM) and replies to the pointer address for scalability requirements and managing the model access control.

### D) Orchestrator:

It realizes and organizes each interaction; it is a manager entity that organizes the federated learning execution steps like a gateway to the blockchain network; in other words, its functionalities are limited and it is a distrust party.

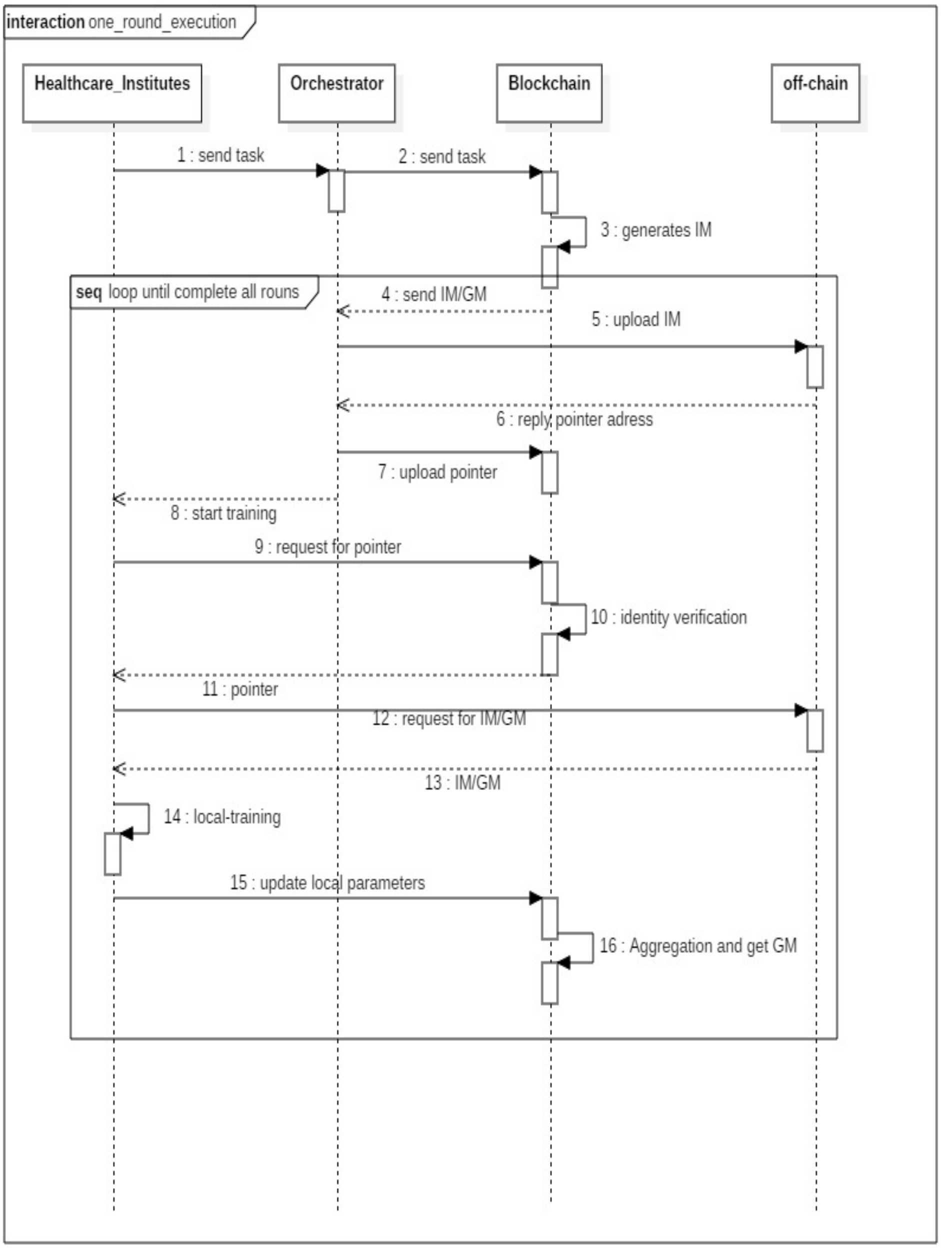


**Figure 2.** MFchain sequenced diagram

### 3.2. MFchain Scenario

In this section, we present the main scenario of the MF chain, where we highlight the interactions between the system components. The system initialization starts by creating the genesis block, which is the first block in the blockchain network, where the peers generate the genesis block that may contain the following attributes: initial model parameters, number of training rounds, public keys of legitimate participants, and hash for assuring model integrity. Figure 2 illustrates that the complete training process of one round can be formulated in the following steps:

1. Healthcare institutes determine and send the training task to the orchestrator after registration, in its turn inquires the Blockchain ❶ for generating the initial model IM (the genesis block). BC replies the IM to Orchestrator ❶, it saves the IM parameters into Off-chain storage and recuperates the pointer address ❷.

2. For managing the access control and allowing only the legitimate participants to obtain the IM, each healthcare institute inquires the BC for the pointer address about the IM ❸. At this time, healthcare institutes ask the off-chain ❹ for the IM by pointer address, knowing that the pointer address is encrypted by the public key of each healthcare institute. Afterward the participant trains the model locally❺ based on its collected clinical raw data, and uploading the result parameters to BC❻.
3. After receiving all notification about completing the local training, the peers of Blockchain generating the GM by using aggregation algorithms ❼, and storing the GM into off-chain through Orchestrator ❽, as well as recuperating the pointer address and the steps are repeated as the IM until completing all the training rounds.

### 3.3. MFchain Preserves Federated learning Privacy

After model generation, the permissioned blockchain is eager to maintain model security and privacy simultaneously. The application of a permissioned blockchain authenticates network members, which can guarantee that only legitimate parties can access the network. For instance, the use of off-chain storage to ensure system scalability makes it difficult to store a model with a large number of parameters. However, we can keep only the address that represents a pointer figure 2 illustrates the sequenced diagram that demonstrates the detailed steps of MFchain, the pointer address is kept on the chain and the only way to recuperate both IM/GM model via grant permission from Blockchain, and after authentication by pseudonym $(pk_{\ i})$ only the legitimate member has the right to inquire for pointer address. Implementation

The proposed approach was implemented under Eclipse using various Java APIs, such as JSON and Fabric SDK. The architecture of the framework is illustrated in Figure 3, where the main components are as follows:

- The Fabric Hyperledger Blockchain is configured for two organizations and one peer node for each organization. The fabric network uses CouchDB as a world-state database and an ordering service. It was built with certificate authority for each organization. Three channels are created for the participants and the Global Model, named respectively “Participants,” “Global Model and “Logs”. Three Fabric smart contracts are deployed using the Go language (one for each channel). The Hyperledger Fabric network used by the proposed method is shown in Figure 4, where every channel is associated with its ledger and smart contract.
- MangoDB: is used as an off-chain storage for the global model, where all data are stored as JSON objects. Queries are specified using NoSQL to interrogate the database and obtain or modify the global model.

## 4. Implementation

The proposed approach was implemented under Eclipse using various Java APIs, such as JSON and Fabric SDK. The architecture of the framework is illustrated in Figure 3, where the main components are as follows:

- The Fabric Hyperledger Blockchain is configured for two organizations and one peer node for each organization. The fabric network uses CouchDB as a world-state database and an ordering service. It was built with certificate authority for each organization. Three channels are created for the participants and the Global Model, named respectively “Participants,” “Global Model and “Logs. “ Three Fabric smart contracts are deployed using the Go language (one for each channel). The Hyperledger

Fabric network used by the proposed method is shown in Figure 4, where every channel is associated with its ledger and smart contract.

- MangoDB: is used as an off-chain storage for the global model, where all data are stored as JSON objects. Queries are specified using NoSQL to interrogate the database and obtain or modify the global model.

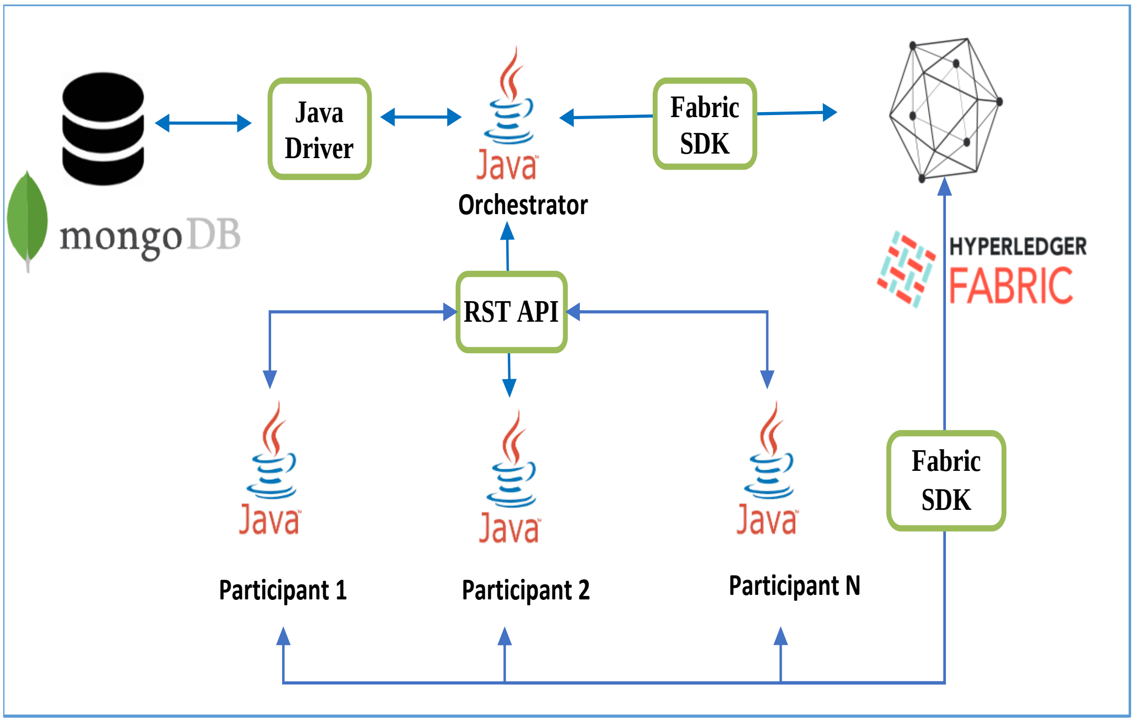


**Figure 3.** Implementation Architecture of MFchain

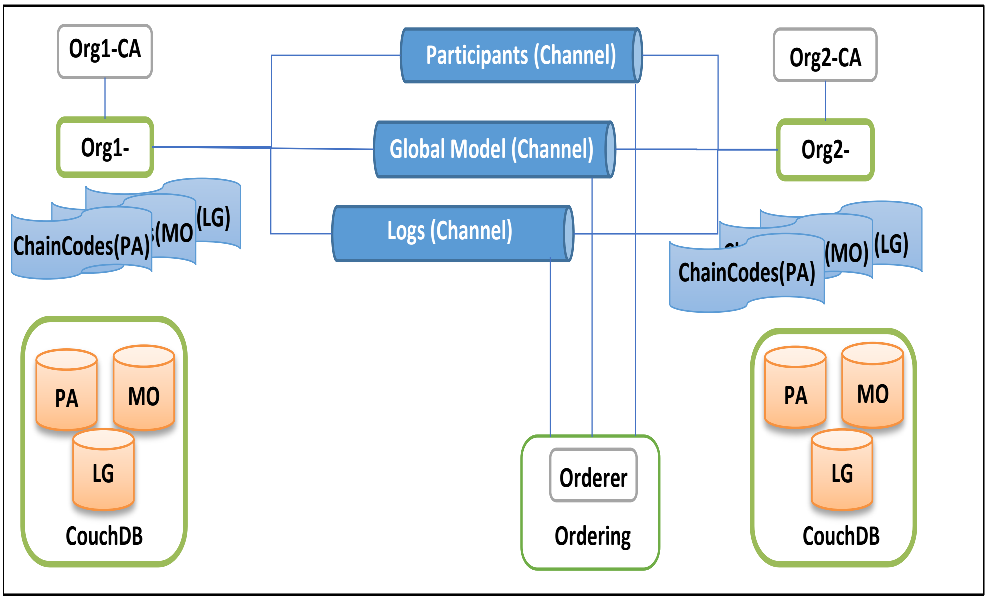


**Figure 4.** Fabric Network Components that are used in MFchain

### 4.1. Fabric Chaincodes

Every peer in HLF has a local database (ledger), which contains all transactions executed by the network via HLF chaincodes. Thus, each peer can have several installed chain codes for a single HLF channel. The distributed ledgers in HLF are updated using smart contracts in demand by external blockchain users. Our work proposes the use of three distributed ledgers, each associated with one smart contract and several peers. These ledgers store critical data about the implementation components, such as participants, Global Model, and operation logs.

The global model chain code uses the Golang structure illustrated in Listing 1. The attribute “Model_hash” calculated by BC peers was used to verify the integrity of the model. Thus, illegal offchain changes were detected.

```
type KerasModel struct {
Model_hash    string `json:"Model_hash"`
ClassName    string `json:"class_name"`
KerasVersion string `json:"keras_version"`
Config struct {
  Layers []struct {
  ClassName string `json:"class_name"`
  Config    struct {
  KernelInitializer struct {
  ClassName string `json:"class_name"`
  Config    struct {
  Distribution string  `json:"distribution"`
  Scale       float64    `json:"scale"`
  Seed        interface{} `json:"seed"`
  Mode         string      `json:"mode"`
                                    } `json:"config"`
                 } `json:"kernel_initializer"`
  Name        string      `json:"name"`    }}}
```

**Listing 1.** The Golang Structure used by the Service Provider Chaincode

### 4.2. MangoDB JSON Struture

Off-chain storage uses the JSON structure given by listing 2 to save the data on the global model. The model hash is not stored in the offchain because it is considered sensitive data that is kept only on the BC.

```
{"class_name": "Sequential",
"keras_version": "2.2.4",
"config": {"layers": [
{"class_name": "Conv2D", "config": {"kernel_initializer": {"class_name": "VarianceScaling",
"config": {"distribution": "uniform", "scale": 1.0, "seed": null, "mode": "fan_avg"}}, "name":
"conv2d_1", "kernel_constraint": null,
"bias_regularizer": null,
"bias_constraint": null, "dtype": "float32", "activation": "relu", "trainable": true, "data_format":
"channels_last", "filters": 64, "padding": "valid"}}]}}
```

**Listing 2.** The JSON Structure used by MangoDB Storage

## 5. Experiment

### 5.1. Expirement Results

During the experiment, our focus was not on enhancing learning, but on the storage side of the global model on both the offchain and onchain sides. We saved a Keras model using these two methods. The first saves the global model in a fabric network using CouchDB as a peer database. The second stores it using the MangoDB server for off-chain storage. The model contains 5,902,151 trainable parameters. In the first experiment, the model size was 3.1 Kilobytes where which was increased for every model update because the BC must maintain the history of all model parameters. In the second experiment, the model size was 20 Kilobytes, which remained unchanged.

### 5.2. Security Analysis

In this subsection, we provide a security analysis of the security and privacy requirements of authentication, authorization, confidentiality, and integrity.

#### 5.2.1. Security

Among the security threats that may face our system is the compromise of the orchestrator. According to the MF chain, the requester must obtain permission from the owner. If the Orchestrator is compromised, it cannot modify anything in the system because it is a distrust entity MFchain has limited its functionalities, it is a blind entity; in other words, it cannot see the model plaintext that the exchanged flow of data has been encrypted. We assume that an external adversary (foreign participant) can use a forged identity to access the system by generating a key pair in order to stole GM parameters; to tackle this problem, the MSP of the permissioned blockchain has provided identities for only legitimate members; in the authentication process, if the public key is not among the list of MSP identities, access is denied. Another assumption is that if the external adversary can violate the identity of a legitimate party, he/she cannot interact with the system without the private key because each participant (healthcare institute) signs his message.

| Method | Permissioned BC | privacy | BC Model Access control | BC scalability |
|---|---|---|---|---|
| El Rifai et al [3] | No | Model | No | No |
| Abou El Houda [4] | Yes | Model | No | No |
| Sun, J et al [5] | Yes | Model | No | No |
| MFchain | Yes | Model+ participant | Yes | Yes |

**Table 1.** Comparison with MFchain and related federated privacy preserving methods

#### 5.2.2. Privacy

The main purpose of MFChain is to maintain federated learning privacy by ensuring scalability. The Global model or local model parameters are encrypted, and only the owner and requester who has the grant can know the plaintext of the model's parameters to preserve the privacy of the IM/GM. (public keys), and the blockchain manages model access using an address pointer. Furthermore, MFchain keeps participants' real identities hidden through the use of pseudonyms.

### 5.2.3. Comparison:

Table 1 presents a comparison between our model and [3], [4], [5] federated learning privacy-preserving approaches, which are very close to our design. They focus on using the blockchain as a trust central server, and most of them have utilized blockchain's storing features to save model weight parameters; in contrast, we integrate permissioned blockchain for managing model access control at the same time using an off-chain storage for scalability issues instead of storing on-chain to obtain an efficient system.

## 6. Conclusion

In this study, we proposed MFchain, which is a new federated learning privacy-preserving framework. MFchain manages model access control via grant permission for the only legitimate participant in blockchain, and for better security and scalability, off-chain storage is supported. MFchain uses a log ledger for auditable history. Our work can be seen as a general framework that can be used to enhance privacy in different fields using secure federated learning frameworks. In future research, we will use two layers of on-chain learning for local learning and the other for global training.

## References

[1] Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., Caro, A.D., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K.A., Sorniotti, A., Stathakopoulou, C., Vukolic, M., Cocco, S.W., & Yellick, J. (2018). Hyperledger fabric: a distributed operating system for permissioned blockchains. Proceedings of the Thirteenth EuroSys Conference.

[2] Dwork, C. (2006). Differential privacy. In Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33 (pp. 1-12). Springer Berlin Heidelberg.

[3] El Rifai, O., Biotteau, M., de Boissezon, X., Megdiche, I., Ravat, F., & Teste, O. (2020). Blockchain-based federated learning in medicine. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18* (pp. 214-224). Springer International Publishing.

[4] Abou El Houda, Z., Hafid, A. S., Khoukhi, L., & Brik, B. (2022). When Collaborative Federated Learning Meets Blockchain to Preserve Privacy in Healthcare. IEEE Transactions on Network Science and Engineering.

[5] Sun, J., Wu, Y., Wang, S., Fu, Y., & Chang, X. (2021). Permissioned blockchain frame for secure federated learning. IEEE Communications Letters, 26(1), 13-17.

[6] Mcmahan H et al (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54: 1273–1282

[7] Yang Q, Liu Y, Chen T, Tong Y (2019) Federated Machine Learning:Concept and Applications. ACM Trans Intell Syst Technol 10(2):1–19

[8] Zyskind, G., Nathan, O., Pentland, A. (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data. 2015 IEEE Security and Privacy Workshops, 180-184.

[9] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-19.

# Intelligent Fault Detection base on machine learning in Autonomous Vehicles : A Hybrid Approach

Amel Sahki[1,2,3], Lokman Touil[3] and Ali Abdelatif Betouil[2,3]

[1] *Laboratoire d'étude et de recherche en instrumentation et en communication d'Annaba (LERICA), Algeria*
[2] *Laboratoire d'informatique et des mathématiques appliquées (LIMA), El Tarf, Algeria*
[3] *Computer Sciences department, Chadli Bendjedid University, El Tarf, Algeria*

## Abstract

This paper presents a hybrid approach combining several machine learning methods for fault detection in autonomous vehicles. The proposed methodology leverages temporal data analysis to capture dynamic fault patterns and robust classification techniques to ensure high diagnostic accuracy. By integrating complementary strengths of machine learning models, this approach effectively handles complex and diverse datasets, addressing challenges such as variability in real-world scenarios and the need for rapid fault identification. Extensive experimentation demonstrates the reliability of the hybrid approach, with significant improvements in fault classification accuracy and robustness. The study also evaluates the computational efficiency of the models, highlighting their suitability for real-time implementation in dynamic environments. Key findings underscore the potential of hybrid methods to enhance predictive maintenance and improve the operational safety of autonomous vehicles.
The paper concludes with insights into the limitations of the current approach and offers perspectives for future research, such as integrating real-time data streams, exploring additional hybrid combinations, and scaling the system for broader automotive applications.

## Keywords

Fault Detection, Classification, Random Forest, SVM, CNN.

## 1. Introduction

Autonomous vehicles rely on a multitude of sensors and control systems to operate safely and efficiently, making **fault detection and classification** essential for maintaining system integrity and ensuring passenger safety [1]. The increasing complexity of these systems has driven the need for advanced classification methods capable of rapidly identifying and categorizing faults, from mechanical and electronic issues to software glitches. Accurate fault classification not only supports **predictive maintenance** but also minimizes system downtime, improving overall vehicle reliability. [2]

There are various types of faults, and it's essential for engineers and developers to understand them. The most common include: **Hardware faults**, which involve issues with physical components, such as sensors or actuators, that can fail or degrade over time; **Software faults**, which stem from errors in the code or algorithms controlling the vehicle, leading to unexpected behaviors; and **Communication faults**, which arise from data-sharing issues between different parts of the system, potentially disrupting operations.[3]

The field of fault diagnosis in autonomous vehicles has evolved significantly, integrating both traditional and advanced methodologies. Traditional approaches include **model-based** which rely on mathematical models of vehicle systems to predict faults based on deviations from expected behavior [4], **signal-based** that analyze sensor data to detect anomalies indicative of faults [5], **knowledge-based methods** that utilize expert knowledge and rules to identify faults, although they often struggle with the complexity of modern vehicles [6], and **Data-driven methods** leverage large datasets and advanced algorithms to identify faults by recognizing patterns and anomalies, though they can face challenges with the variability and complexity of real-world scenarios [7].

Traditional diagnostic methods often struggle to manage the complexities of modern vehicles. Several machine learning methods have been applied to the classification of faults in autonomous vehicles, among these methods **Support Vector Machines (SVM)**, **Random Forest (RF)** and **k-Nearest Neighbor (KNN)**.

In contrast, deep learning techniques have become valuable tools for fault identification, utilizing large datasets to improve diagnostic accuracy [8].

Among the classification methods applied in this field, including deep neural networks like convolutional (CNN) and recurrent neural networks (RNNs), which are used to detect fault patterns [9]. The Deep Symptoms-Based Model (Deep-SBM) analyzes symptoms to predict faults and has shown improved performance over traditional methods [10]. Additionally, dynamic model learning allows deep learning methods to identify multiple fault types in real-time by analyzing both input and output signals [11]. Intelligent fault classification systems further enhance diagnostics by automating classification based on driving data and operational conditions [12], while IoT-enabled real-time data utilization improves system responsiveness for online fault prediction [13].

While deep learning offers significant advancements in fault diagnosis, traditional methods still play a role, particularly in simpler systems. The integration of both approaches may yield the most robust solutions for future autonomous vehicle diagnostics [14]. **Hybrid models**, combining two or more of these techniques, are increasingly employed to harness their complementary strengths for instance. In our work, we present a hybrid approach combining **CNN with Random Forest** to capitalize on temporal data analysis while maintaining robust classification.

This paper provides an overview of these classification methods, examining their applications in autonomous vehicle fault detection, with an emphasis on model accuracy, robustness, and suitability for real-time fault classification in dynamic environments. By exploring the strengths and limitations of each method, we aim to highlight strategies that enhance fault detection systems and support the safe operation of autonomous vehicles.

## 2. Related work

Autonomous vehicles (AVs) rely on a complex network of sensors and control systems in order to navigate and make decisions, so for the past 30 years many algorithms have been proposed for fault detection due to the need for an accurate fault detection system for this type of vehicles have become a necessity. The use of machine learning and deep learning algorithms aims to enhance the accuracy of these systems using different types of data like Nasim, Fawad, et al (2023) used sound as data and processed it with ELM to identify the different types of engine faults [15], Pavlopoulos et al (2024). proposed a machine learning and natural language processing-based automotive fault diagnosis. Their model (Transformers) focused on classifying text-based fault reports from multilingual datasets with accuracy rates above 80% for common fault categories. It enabled fast and accurate fault detection through the use of a multilingual Transformer-based approach to enhance automotive troubleshooting [16].

### 2.1. Machine Learning Approaches

Machine learning approaches have been widely used for fault detection due to their capabilities to identify patterns and anomalies and their adaptability to evolving fault conditions.

In Biddle and Fallah (2021), a new fault detection, isolation, and prediction approach was developed for multi-sensor systems in autonomous vehicle controllers using SVM. It investigates the extension for multi-fault situations by conflating sensor signals for detection purposes using a single-class SVM model, with better applicability in real time. Sensor-specific SVMs were utilized afterward to identify the faulty component in fault isolation, while multi-class SVMs classified the type of fault. Their architectures were verified to yield detection, isolation, and identification accuracies of 94.94%, 97.42%, and 97.01%, respectively, using simulated driving data. This module projects the advancement in sensor degrading, thus assisting in proactive vehicle health management [17].

In the case of Mishra et al., RF was applied to fault detection and diagnosis of electric vehicles. The model develops an ensemble of decision trees that underwent training on features and random subsets of data; therefore, it boosts model accuracy and protects from overfitting. In their case, the RF classifier showed outstanding training and testing performance: 99.85% and 97.44%, respectively. This good performance in terms of fault identification and diagnosis underlines the capability of the RF algorithm to handle complex, diverse patterns in datasets in an efficient way, which can be considered a very valuable tool in enhancing EV reliability and proactive maintenance [18].

Table 1 presents a comparative overview of various fault detection models used in autonomous vehicle diagnostics. The table highlights the type of model, the dataset employed, and the achieved accuracy, providing insights into the performance and applicability of each approach. This comparison underscores the potential of advanced machine learning methods, such as deep neural networks and hybrid techniques, to achieve high diagnostic accuracy, while also identifying areas for further improvement.

**Table 1**
Comparative Performance of Fault Detection Models

| Reference | Model | Dataset | Accuracy |
|---|---|---|---|
| [15] | DNN | Simulated vehicle data (500 normal, 500 faulty samples) | 100% |
| [16] | SVM | Multi-sensor systems data for autonomous vehicle controllers | Detection: 94.94%, Isolation: 97.42%, Identification: 97.01% |
| [17] | Random Forest | Simulated EV data (healthy & faulty) | 97.44% |
| [18] | Multilingual Transformer | Multilingual automotive fault reports dataset (text-based) | 80%+ for high-frequency classes |
| [19] | ELM | Real-world vehicle sounds from Ford and Toyota models | 92.17% |

## 2.2. Deep Learning Approaches

The previous ML models are called shallow learning models despite their great achievements but they struggle in analyzing complex data with no prior knowledge, that's why deep learning have been developed to tackle this problem.

Ren et al. (2024) developed a comprehensive fault detection framework for autonomous vehicles. The process starts with building a mathematical model of the vehicle based on sensor data from the input signals applied to the motors and the heading angle. This model allows them to simulate mechanical and electrical faults, generating 500 training and 500 testing samples (500 normal, 500 faulty). Since the input signals (voltage applied to the motors) are 1D, they converted them into 2D time-frequency images using a wavelet transform. These images were then fed into a 10-layer deep neural network (DNN), which learned to detect faults. The model was evaluated on the test data, achieving 100% accuracy, demonstrating the high effectiveness of their approach [19].

# 3. Methodology

## 3.1. Dataset description

The dataset we previously used represents a collection of vehicle engine failure data, designed for classification and fault detection. Here is a detailed description of each aspect of this dataset:

### 3.1.1. Data Collection

Data was collected from sensors installed on test vehicles, covering various engine faults, including misfires, compression leaks, and mechanical failures. The data includes vibrational, acoustic, and thermal signals, obtained during driving cycles that simulate real conditions (acceleration, deceleration, cruising speed). Initial data volume was insufficient for certain fault classes, prompting the use of data augmentation.

### 3.1.2. Data Preparation

Before running the classification models, rigorous data preprocessing was performed:

- **Signal Filtering:** Low-pass and high-pass filters were applied to remove background noise and retain characteristic frequencies.
- **Feature Extraction:** Transformation of time-domain signals to the frequency domain (FFT, Fast Fourier Transform), along with statistical feature calculations such as mean, standard deviation, and dominant frequencies.
- **Normalization:** Extracted features were normalized to reduce the impact of amplitude variations and facilitate model convergence.

### 3.1.3. Dataset Structure Title marks

The dataset consists of rows, where each row represents an observation or record of a specific engine, with data on its sensors as well as information about the fault or observed behavior. The columns include sensors measuring different parameters and other contextual information for in-depth analysis of engine failures.

**Table 2**
Head-Lines and Variable Descriptions.

| Head-lines | Type | Description |
|---|---|---|
| Recording Date | Date | The date when the observation was recorded, allowing for tracking of trends and timing of failures. |
| Failure Type | Categorical | The type of engine failure or fault (e.g., misfire, abnormal noise, overheating, etc.). |
| Problem Description | Text | A brief description of the detected issue. For example: "Abnormal noise during operation." |
| Sensor 1 | Numeric | Represents engine vibrations in standardized units. |
| Sensor 2 | Numeric | Indicates the engine temperature. |
| Sensor 3 | Numeric | Shows the oil pressure. |
| Sensor 4 | Numeric | Measures the instant fuel consumption. |
| Sensor 5 | Numeric | Measures the battery voltage. |
| Sensor 6 | Numeric | Measures CO₂ emissions in g/km. |
| Sensor 7 | Numeric | Measures engine speed (RPM). |

### 3.1.4. Data Normalization

Standard normalization involves transforming each feature so that it has a mean of 0 and a standard deviation of 1. It is defined by the formula:

$$Z = \frac{(x - \mu)}{\sigma} \quad (1)$$

where:

- Z is the normalized value,
- $x$ is the original value of the feature,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

The values in each column will thus be distributed around zero with a certain spread defined by the standard deviation. If the initial values of a feature are below the mean, their transformation will yield negative values. This does not indicate an error but simply that the values are below the mean.

To use a Support Vector Machine (SVM) for the classification of our normalized dataset, we will follow these steps:

1. Load the data and separate the features ($x$) from the target variable (y).
2. Split the data into training and testing sets.
3. Train an SVM model on the training data.
4. Predict the classes on the test data.
5. Calculate accuracy and display the confusion matrix.

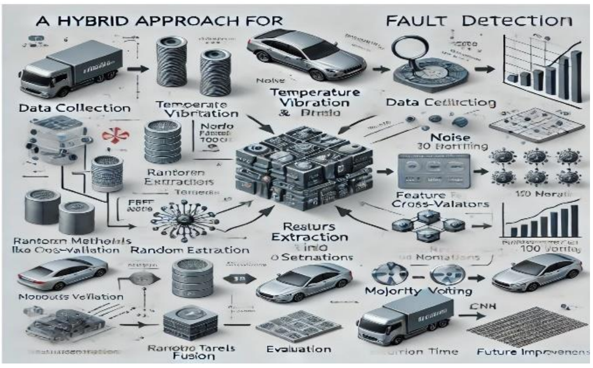


**Figure 1:** Flowchart of the Hybrid Approach for Fault Detection in Autonomous Vehicles

### 3.1.5. Classes of Failure Types

- **Misfire:** Issues within the engine's ignition system.
- **Abnormal Noise:** Unusual vibrations or noise during operation.
- **Overheating:** Engine temperature exceeding normal values.
- **Lubrication Fault:** Issues related to oil pressure or insufficient lubrication.
- **Battery Discharge:** Insufficient battery charge, affecting startup or performance.

## 3.2. Proposed methods

For classification, several models were used, each with optimized parameters and specific justifications:

1. **Support Vector Machines (SVM):** An RBF (Radial Basis Function) kernel was used to better capture non-linear relationships in the data. The regularization parameter C was adjusted to balance complexity and overfitting.
2. **Random Forest:** This model consisted of 100 trees to achieve a good trade-off between accuracy and computation time. Decision trees are constructed with random samples of features, making the model more resistant to overfitting and capable of capturing data variance.

3. **SVM - Random Forest Combination:** Predictions from both models are combined through majority voting to obtain a final classification. This approach leverages the SVM's ability to detect fine separations between classes and the Random Forest's robustness to noisy data.

4. **Convolutional Neural Network (CNN):** A CNN model was designed with 1D convolutional layers, suitable for time-series signals from engine sensors. The model consists of three convolutional layers with filter sizes of 16, 32, and 64 to capture features at different granularities, followed by pooling layers for dimensionality reduction. The final layer uses a softmax activation function for multi-class classification.

5. **Data Augmentation:** Specific data augmentation techniques for time-series signals were applied, including:

- **Gaussian Noise Addition:** To simulate realistic signal variations.
- **Time Shifting:** Modifying signal start points to simulate shifted engine cycles.
- **Frequency Transformation:** Slight frequency variations to simulate different engine regimes.

## 4. Results and discussion

To evaluate the effectiveness of each approach, models were trained and tested using 10-fold cross-validation. Details of the experiments include:

- **SVM and Random Forest Training:** These models were initially trained on raw data, then on augmented data. The SVM model was fine-tuned with grid search to optimize the C parameter and gamma kernel parameter. The Random Forest was tested with various numbers of trees (from 50 to 150) to assess the impact on accuracy and processing time.
- **CNN Training:** The CNN was trained on augmented data, using a cross-entropy loss function for multi-class classification and Adam optimization with an initial learning rate of 0.001. Regularization techniques, such as dropout, were added to reduce overfitting, particularly given the model's complexity and the relatively small dataset size.

**Table 3**
Results

| Model | Accuracy | Training Time | Classification Time |
|---|---|---|---|
| SVM | 73 % | 10 min | 1.8 s |
| Random Forest | 75 % | 8 min | 1.7 s |
| SVM-RF Combination | 79 % | 12 min | 2.2 s |
| CNN | 89% | 23 min | 3.8 s |
| CNN (Data Augmentation) | 93% | 29 min | 4.3s |

The table presents the performance of different classification models in terms of accuracy, training time, and classification time. The SVM model, with an accuracy of 73%, is fast for classification (1.8 s) but less accurate. The Random Forest, with 75% accuracy and an 8-minute training time, offers a good balance by being quick in both training and classification. The SVM-RF combination slightly improves accuracy to 79% but has a slightly longer classification time (2.2 s). The CNN (Convolutional Neural Network) model achieves a much higher accuracy of 89%, at the cost of a longer training time (23 min) and a classification time of 3.8 s. Finally, the CNN with data augmentation provides the best accuracy (93%) but requires the longest training and classification times, at 29 minutes and 4.3 s, respectively. In conclusion, while the CNN with data augmentation is the most accurate, it is also the most time-consuming, whereas the SVM and Random Forest models remain good options for balancing accuracy and speed.

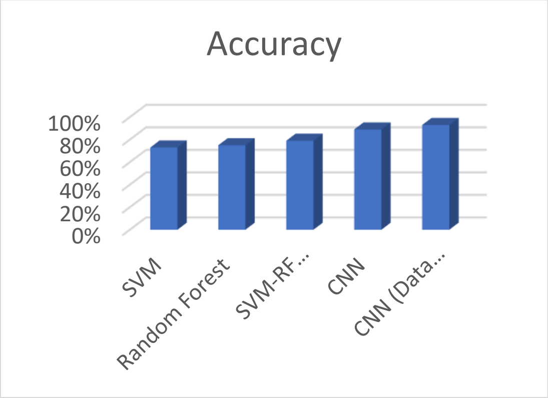


**Figure 2:** Comparison of Model Accuracy for Fault Detection

Figure 2 illustrates the accuracy comparison of various fault detection models for autonomous vehicles. The models include Support Vector Machine (SVM), Random Forest, a hybrid of SVM and Random Forest (SVM-RF), Convolutional Neural Network

(CNN), and CNN with Data Augmentation. SVM and Random Forest show strong performance individually, with SVM benefiting from its RBF kernel for capturing non-linear patterns and Random Forest leveraging decision tree ensembles. The hybrid SVM-RF model further enhances accuracy by combining the strengths of both methods. CNN also delivers high accuracy, especially for complex data, and the use of data augmentation with CNN improves generalization, leading to the highest accuracy. This comparison underscores the potential of hybrid models and data augmentation in improving diagnostic reliability for fault detection in autonomous vehicles.

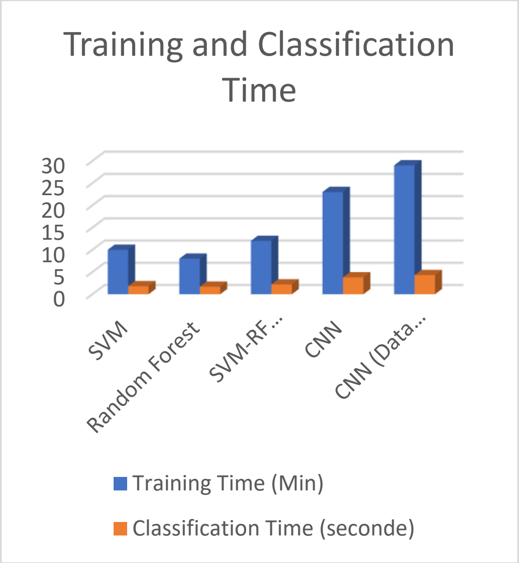


**Figure 3:** Flowchart of Comparison of Training and Classification Time for Fault Detection Models

Figure 3 compares the training and classification times for different fault detection models used in autonomous vehicles. The chart distinguishes between the time required for training (in minutes) and classification (in seconds). The **SVM** and **Random Forest** models exhibit relatively low training times, with Random Forest slightly higher due to its ensemble nature. However, their classification times remain fast. In contrast, the **SVM-RF hybrid** model has a slightly longer training time, reflecting the added complexity of combining both models, but still maintains efficient classification times. The **CNN** model shows significantly higher training times, as expected due to its deep learning architecture, but the classification time remains competitive. The **CNN with Data Augmentation** takes the longest training time, as expected, due to the additional data preprocessing step, but its classification time does not substantially increase, indicating that the augmentation process enhances model generalization without a major trade-off in inference time. This comparison highlights the trade-offs between model complexity, training duration, and real-time performance for fault detection systems in autonomous vehicles.

## 5. Conclusion

This study underscores the potential of hybrid approaches in advancing fault detection capabilities for autonomous vehicles. By leveraging the strengths of Convolutional Neural Networks (CNN) in conjunction with data augmentation techniques, our method achieves significant improvements in classification accuracy and robustness. These advancements demonstrate the effectiveness of combining advanced machine learning techniques to address the complexities and variability inherent in fault diagnosis for autonomous systems.

The proposed approach not only enhances the reliability of fault detection systems but also contributes to predictive maintenance, reducing vehicle downtime and improving operational safety. However, there is room for further enhancement. Future research will aim to explore additional hybrid combinations, such as integrating other machine learning models or deep learning architectures, to optimize performance across various datasets. Moreover, integrating real-time data streams will be a key focus, allowing the system to operate dynamically and adapt to evolving fault conditions. Testing the approach on larger, more diverse datasets, including real-world scenarios, will be essential to validate its scalability and applicability across different types of autonomous vehicles. By addressing these aspects, we aim to contribute to the development of more robust and reliable diagnostic systems for the next generation of autonomous vehicles.

## References

[1] Bathla, Gourav, et al. "Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities." Mobile Information Systems 2022.1 (2022): 7632892.

[2] Abid, Anam, Muhammad Tahir Khan, and Javaid Iqbal. "A review on fault detection and diagnosis techniques: basics and beyond." Artificial Intelligence Review 54.5 (2021): 3639-3664.

[3] Hall, Tracy, et al. "A systematic literature review on fault prediction performance in software

engineering." IEEE Transactions on Software Engineering 38.6 (2011): 1276-1304.

[4] Isermann, Rolf. "Model-based fault-detection and diagnosis–status and applications." Annual Reviews in control 29.1 (2005): 71-85.

[5] Gao, Zhiwei, Carlo Cecati, and Steven X. Ding. "A survey of fault diagnosis and fault-tolerant techniques—Part I: Fault diagnosis with model-based and signal-based approaches." IEEE transactions on industrial electronics 62.6 (2015): 3757-3767.

[6] Li, Weijun, et al. "Process fault diagnosis with model-and knowledge-based approaches: Advances and opportunities." Control Engineering Practice 105 (2020): 104637.

[7] Jieyang, Peng, et al. "A systematic review of data-driven approaches to fault diagnosis and early warning." Journal of Intelligent Manufacturing 34.8 (2023): 3277-3304.

[8] Zhu, Zhiqin, et al. "A review of the application of deep learning in intelligent fault diagnosis of rotating machinery." Measurement 206 (2023): 112346.

[9] Mnyanghwalo, Daudi, et al. "Deep learning approaches for fault detection and classifications in the electrical secondary distribution network: Methods comparison and recurrent neural network accuracy comparison." Cogent Engineering 7.1 (2020): 1857500.

[10] Arena, Fabio, et al. "Predictive maintenance in the automotive sector: A literature review." Mathematical and Computational Applications 27.1 (2021): 2.

[11] Abboush, Mohammad, et al. "Intelligent fault detection and classification based on hybrid deep learning methods for hardware-in-the-loop test of automotive software systems." Sensors 22.11 (2022): 4066.

[12] Hossain, Md Naeem, Md Mustafizur Rahman, and Devarajan Ramasamy. "Artificial Intelligence-Driven Vehicle Fault Diagnosis to Revolutionize Automotive Maintenance: A Review." CMES-Computer Modeling in Engineering & Sciences 141.2 (2024).

[13] Mudia, Halim. "Utilization AI for Predictive Maintenance in IoT-Enabled Industrial Systems." Journal of Artificial Intelligence and Development 2.2 (2023): 47-51.

[14] Zhao, Xiangmo, et al. "Potential sources of sensor data anomalies for autonomous vehicles: An overview from road vehicle safety perspective." Expert Systems with Applications 236 (2024): 121358.

[15] Nasim, Fawad, et al. "Intelligent Sound-Based Early Fault Detection System for Vehicles." Computer Systems Science & Engineering 46.3 (2023).

[16] Pavlopoulos, John, et al. "Automotive fault nowcasting with machine learning and natural language processing." Machine Learning 113.2 (2024): 843-861.

[17] Biddle, Liam, and Saber Fallah. "A novel fault detection, identification and prediction approach for autonomous vehicle controllers using SVM." *Automotive Innovation* 4.3 (2021): 301-314.

[18] Mishra, Debani Prasad, et al. "Fault detection and diagnosis of electric vehicles using artificial intelligence." *International Journal of Applied* 13.3 (2024): 653-660.

[19] Ren, Jing, et al. "A deep learning method for fault detection of autonomous vehicles." *14th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2019.

# Intelligent Industrial Process Monitoring: SSAE for Quality Assurance

Wafa Bougheloum [1];Mounir Bekaik [1], Saliha Maarouf [2]; Albdelhamid Ksentini [2]

[1] *dept. of electronic, Badji Mokhtar Annaba University,* Laboratory Automatic and Signals Annaba, Annaba, Algeria

[2] *dept. of electrotechnical, Badji Mokhtar Annaba University,* Laboratory of Electrotechnical, Annaba, Algeria

**Abstract**— Traditional process monitoring techniques, such as Principal Component Analysis (PCA), often rely on the assumption of Gaussian-distributed and linearly correlated process data. However, these assumptions frequently do not hold true in real-world industrial scenarios due to the inherent non-linearity of many processes. This paper introduces a novel multivariate statistical process monitoring framework utilizing Stacked Sparse Autoencoders (SSAE) to reconstruct faulty sensor data.

To effectively monitor process performance, a Squared Prediction Error (SPE) index and an adaptive non-parametric confidence limit derived from kernel density estimation (KDE) are employed. Additionally, an enhanced sensor validity index (SVI), grounded in the reconstruction principle, is proposed to identify defective sensors. Experimental results, encompassing both synthetic and real-world data from a drinking-water treatment plant, demonstrate the efficacy of the proposed scheme and its ability to accurately detect and isolate sensor failures.

**Keywords**— Process monitoring, Multivariate statistical process control, Stacked Sparse Autoencoder (SSAE), Anomaly detection, Data reconstruction

## 1. Introduction

The growing complexity of industrial processes demands advanced monitoring techniques capable of managing large-scale [1], nonlinear, and dynamic systems. Autonomous systems, integrating machine learning and artificial intelligence [2], are transforming the landscape by enabling intelligent, adaptive solutions. This work leverages Stacked Sparse Autoencoders (SSAE) [3] to improve anomaly detection and fault isolation, addressing key challenges faced by traditional methods such as linearity assumptions and limited robustness to noise.

The proposed framework offers a significant step forward in process monitoring by combining state-of-the-art deep learning techniques with adaptive thresholds and reconstruction-based diagnostics [4]. These advancements demonstrate the potential for intelligent systems to enhance industrial reliability, operational efficiency, and decision-making under complex conditions. The results, validated on both synthetic and real-world datasets, highlight the effectiveness of SSAE in providing accurate, robust, and scalable solutions for industrial anomaly detection and fault isolation.

## 2. Stacked Sparse Autoencoders (SSAE)

Deep Learning has demonstrated remarkable success in various tasks, including image processing and visual analysis. While its application in process control is still emerging, deep neural network models offer valuable feature extraction capabilities through hidden layers. Consider an input vector $X_i = \{1, 2, 3, ..., N\}$. In the encoding layer, this vector is transformed into a hidden representation hi using the function:

$$h_i = f(x_i) = sigm(W_1 x + b_1) \qquad (1)$$

Where $W_1$ and $b_1$ are respectively the weight and the bias between the input layer and the hidden part and $sigm(x)$ is a sigmoid function.

In the decoding layer, $h_i$ is mapped to the output denoted by
$x\hat{}$. Where we use the activation function shown as follows:

$$x\hat{}_i = g(h_i) = sigm(W_2 h + b_2) \qquad (2)$$

Where $W_2$ and $b_2$ are respectively the weight and the bias be- tween the hidden part and the output layer ($x\hat{}$).

A Stacked Sparse Autoencoder (SSAE) [10] is a neural network consisting of multiple layers of sparse autoencoders in which the outputs of each layer are wired to the inputs of the successive layer. In this way, the training of multiple sparse autoencoders (figure1) is completed.

SSAE is a network with a sparsity penalty applied to the hidden layer. The goal is to train the network to predict its output (estimation of the input) as close as possible to its input. This is achieved by optimizing the cost function defined by:

$$J = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}\|\hat{x}_i - x_i\|^2\right) + \frac{\lambda}{2}\sum_{i=1}^{N}\|W_i\|^2 + \beta\sum_{j=1}^{m} KL(\rho\|\,\hat{\rho}_j) \quad (3)$$

Where *m* is the number of the hidden node. $\lambda$ and $\beta$ are the coefficient that determine the weight decay and the sparsity penalty terms, respectively.

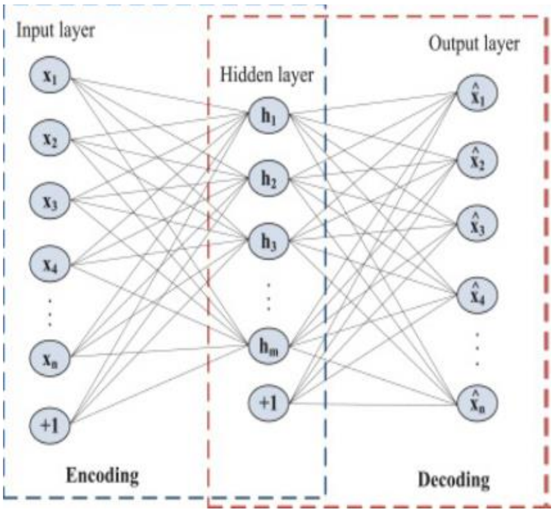


Fig 1. Sparse Autoencoder

In the equation (3), the first term represents the reconstruction error, the second is the regularization term and the last is Sparsity Penalty term, where $KL(\rho\,\|\,\rho\hat{}_i)$ is the Kullback-Leibler divergence, it is used to calculate the difference between $\rho$ and $\rho\hat{}_i$, where $\rho$ and $\rho\hat{}_i$ are the constraint used during learning. The back propagation algorithm is used to minimize the cost function and to find the appropriate parameters $W_1$, $W_2$, $b_1$, $b_2$.

## 2.1 COMPARISON WITH PRINCIPAN COMPONENT ANALYSIS (PCA)

Traditional methods like Principal Component Analysis (PCA) have been widely used for anomaly detection and fault isolation in industrial process monitoring. However, PCA's reliance on linear data relationships and Gaussian distribution assumptions often leads to suboptimal performance in nonlinear, complex systems. In contrast, Stacked Sparse Autoencoders (SSAE) leverage deep learning's ability to capture intricate nonlinear patterns, providing superior detection and reconstruction capabilities.

Key advantages of SSAE over PCA include:

- Improved detection accuracy in scenarios involving nonlinear interactions between variables.
- Lower false alarm rates, especially in noisy environments.
- Enhanced capability for real-time processing due to efficient feature extraction layers.

## 3. ANOMALY DETECTION

Several anomaly detection techniques have been proposed to diagnose the underlying causes of malfunctions. One common approach involves residual analysis, which compares the measured data with estimated data. The Squared Prediction Error (SPE) is a widely used metric for this purpose, defined as:

$$Q = SPE = \sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \quad (4)$$

Where *N* is the number of samples.

### *3.1 Control Limits*

*Parametric $\chi^2$ Distribution ($\delta^2$ ):* The control limit is the process variation that indicates when the process is out of control. The system is considered in its normal δperating conditions if $SPE \leq \delta^2$ . On the other hand, if $SPE > \delta^2$ , the system is considered defective , where $\delta\alpha$ is specified for the SPE control limit [5], which can be calculated using a weighted $\chi^2$ distribution:

$$\delta = g\chi^2_{h,\alpha} \quad g = \frac{v}{2m} \quad h = \frac{2m^2}{v} \quad (5)$$

Where *m* and *v* are respectively the estimated mean and variance of SPE.

### *3.2 Adaptive Threshold Based Kernel Density Estimation K-means Clustering (AUCLKDE K-means):*

The Kernel density estimation (KDE) is a powerful technique for nonparametrically estimating the probability density function of a random variable at any point in its support. Given a sample matrix with n variables and m samples, the KDE of the density function f(x) at any point x is defined as follows:

$$f(x) = \frac{1}{mh}\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \qquad (6)$$

Where h is the bandwidth parameter and K is a kernel function that integrates to one and has zero mean.

Clustering aims to partition data into disjoint subsets called clusters, such that points within the same cluster are more similar to each other than points in different clusters. Among various classification methods, partition clustering, such as K-means, is widely used [6].

K-means is a simple unsupervised learning algorithm that partitions a dataset into *k* predefined clusters, minimizing the sum of squared distances between each data point and the centroid of its cluster. The objective is to find the *k* centroids that minimize the following objective function:

$$J = \sum_{j=1}^{k}\sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2 \qquad (7)$$

## 4. FAULT IDENTIFICATION

### *4.1 Contribution Plots*

There are several methods of faults identification. For this purpose, contribution plots can be used. The contribution of variable *j* to the *Q* statistic is calculated as follows:

$$C_{ijk}^{Q} = e_{ijk}^{2} \qquad (8)$$

Where $e = (x_i - x\hat{}_i)$.

### *4.1.1 Nonlinear Reconstruction Principle*

*To be able to reconstruct the faulty data, it is necessary to determine the fault in a unique way (figure 2).*

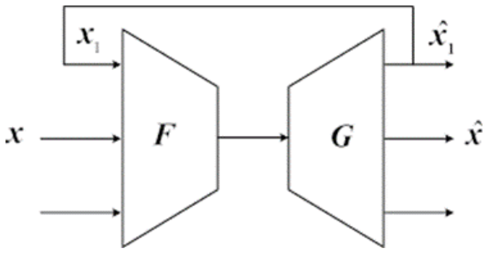


Fig. 2: Reconstruction Principle

The approach consists in predicting the measurement xˆj of the process, by replacing the jth process variable by the predicted one and repeating the operation until convergence of the algorithm as follow:

$$\tilde{x}_i = \xi_j^T G(F(x_j)) \qquad (9)$$

Where $x\tilde{}_i = (x_1, x_2, ..., x\hat{}_j, ..., x_m)$, $\xi^T$ is the $j^{th}$ column of the identity matrix.

### *4.2 Sensor Validity Index (SVI)*

It is the measure of sensor performance where standard range should exist regardless of the number of principal components of the disturbances or faults [7], it is defined as follows:

$$\eta_j^2(k) = \frac{SPE_j(k)}{SPE(k)} \qquad (10)$$

Where SPE is the quadratic global prediction error computed before reconstruction and SP Ej is the jth quadratic prediction error computed after reconstruction [8].

The validity index of a faulty sensor must converge towards zero.

## 5. CASE STUDIES

### *5.1 Synthetic data*

We use dataset [9] containing three variables where *t* is uniformly distributed in the interval [−1, 1]; $\varepsilon_i$ denotes the Gaussian white noises with zero means and standard deviation of 0.01 and 1000 samples collected to build SSAE model.

$$\begin{aligned} x1 &= t2 + 0.3 \sin(2\pi t) + \varepsilon 1 \\ x2 &= t + \varepsilon 2 \qquad (11) \\ x3 &= t3 + t + 1 + \varepsilon 3 \end{aligned}$$

After creating the model, we check the evolution of SPE under normal conditions, where statistical and adaptive threshold is calculated. The result is shown in fig.3. For the clarity of the results, we will use a window from 150 to 250 samples.

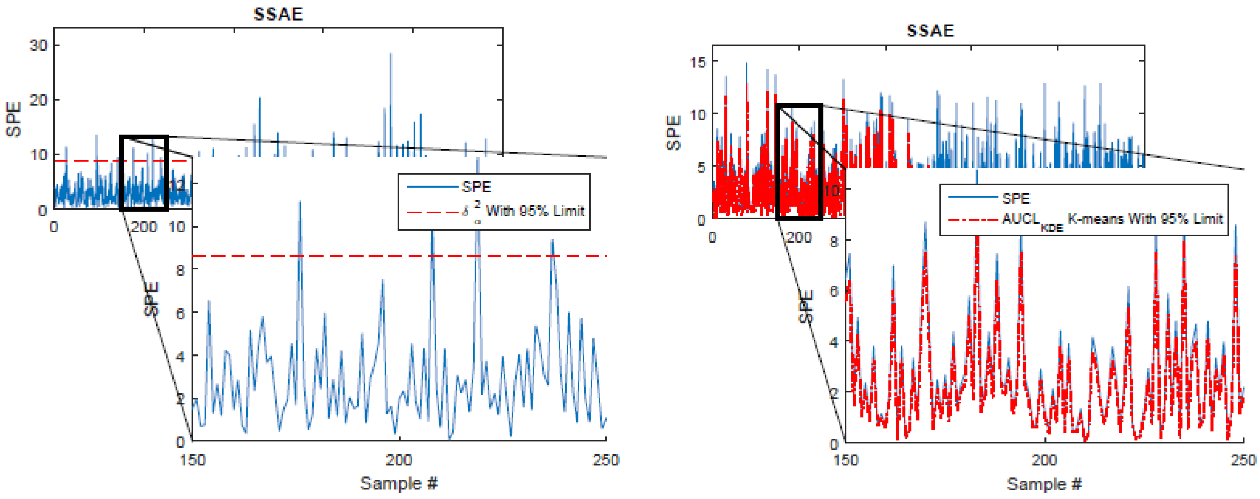


Fig.3a: Statistical threshold

Fig.3b: AUCL threshold using K-means clustering

Fig.3. SPE: data in normal state

We simulate a fault in one of the sensors and we notice the evolution of SPE with the two thresholds as it is illustrated in fig.4.

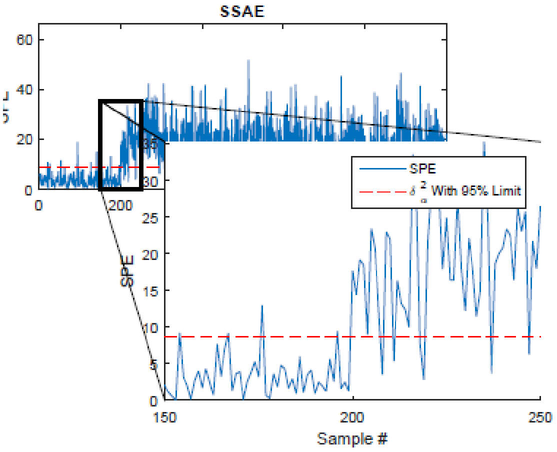


Fig.4a: Statistical threshold

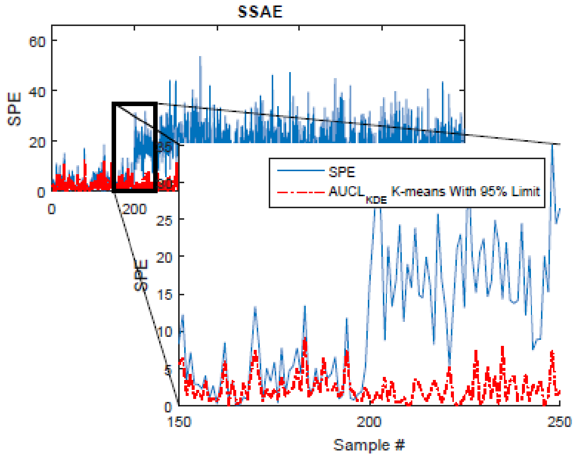
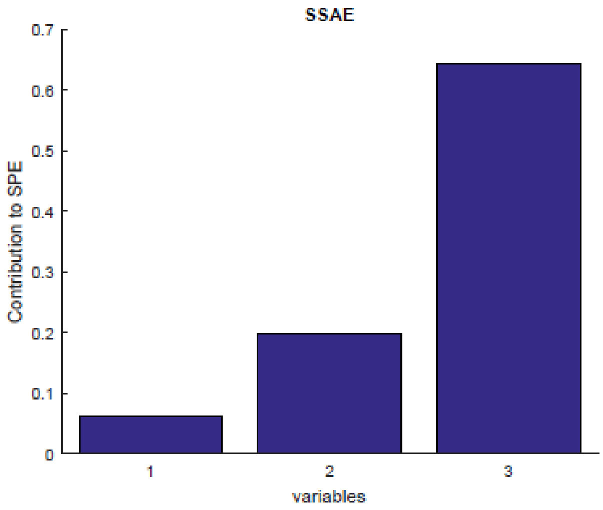


Fig.4b: AUCL threshold using K-means clustering

Fig.4. SPE: data in faulty state

By examining the figures in normal and faulty state, we can see a false alarm in our data caused generally by outlier measures. The SSAE model is able to detect the fault, the detection was at 200 sample. Also, we can identify which sensor is faulty by using contribution plots as it is shown in fig.5. Where the faulty sensor is the third one.



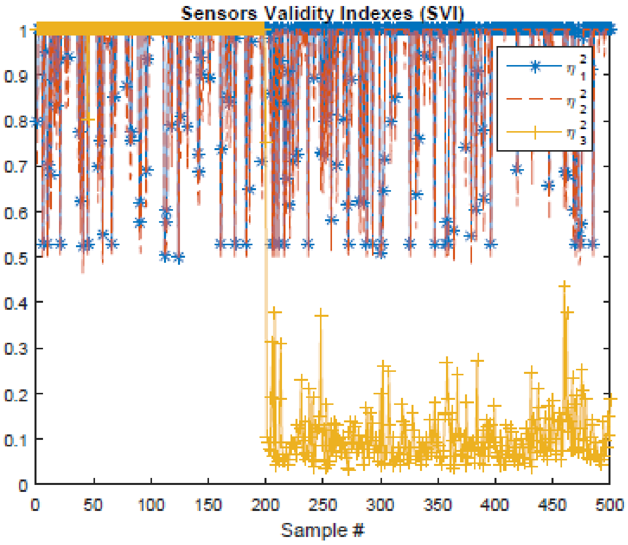


Fig.5. Fault isolation using normalized contribution plots (fault in the 3rd sensor)

Fig.6. Fault identification using Sensor Validity Index SVI (Fault in the 3rd sensor)

There is an other method of faulty sensor localization using SVI, who leads the process to its normal operating conditions as shown in fig.6.

### *5.2 Application On A Water Treatment Plant*

Traditional drinking water treatment that has surface water
usually include four important processes: flocculation, sedimentation, filtration and disinfection, as shown in fig.7.
Adding chemicals to water may be the most important process in the surface treatment plant. The main function of the unit is chemical coagulation, in which chemicals, usually aluminum or iron salts, are added into the water for the purpose of producing flocs of colloidal particles and deposition of other contaminants.
In our study, the model inputs consist of raw water parameters, while the model output is the best dose of

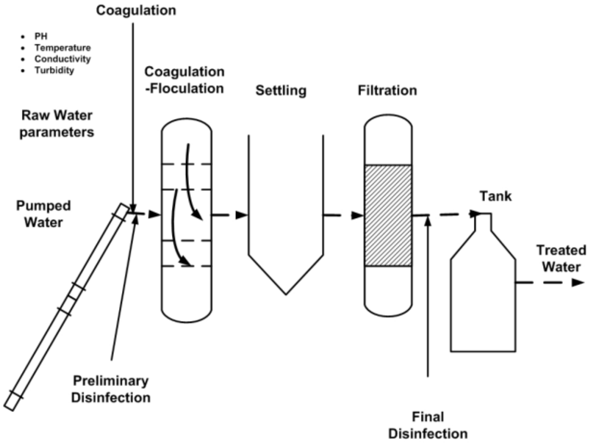


Fig. 7: The typical drinking water processing units

coagulant to achieve the required quality of treated water. This section focuses on the development of self-coagulation control based on the total water parameters to calculate the required dose. The plant we studied is the purification of water in Oued el Othmania. It is responsible for the distribution of drinking water to many citizens at and around Constantine (Algeria) [10].
The data used in this study contains: raw water parameters
and treated water parameters: Turbidity, Temperature, PH and O2, so in total we have eight parameters implying that we have 8 sensors to monitor. The measurements are sampled by a data acquisition unit (SCADA system) covering a period of 356 days, respecting different periods. We use this database to construct and develop the SSAE model.
The result obtained for the actual data for SPE in normal
conditions with the two types of thresholds is illustrated in the fig.8. For the clarity of the results, we will use a window from 150 to 250 samples.

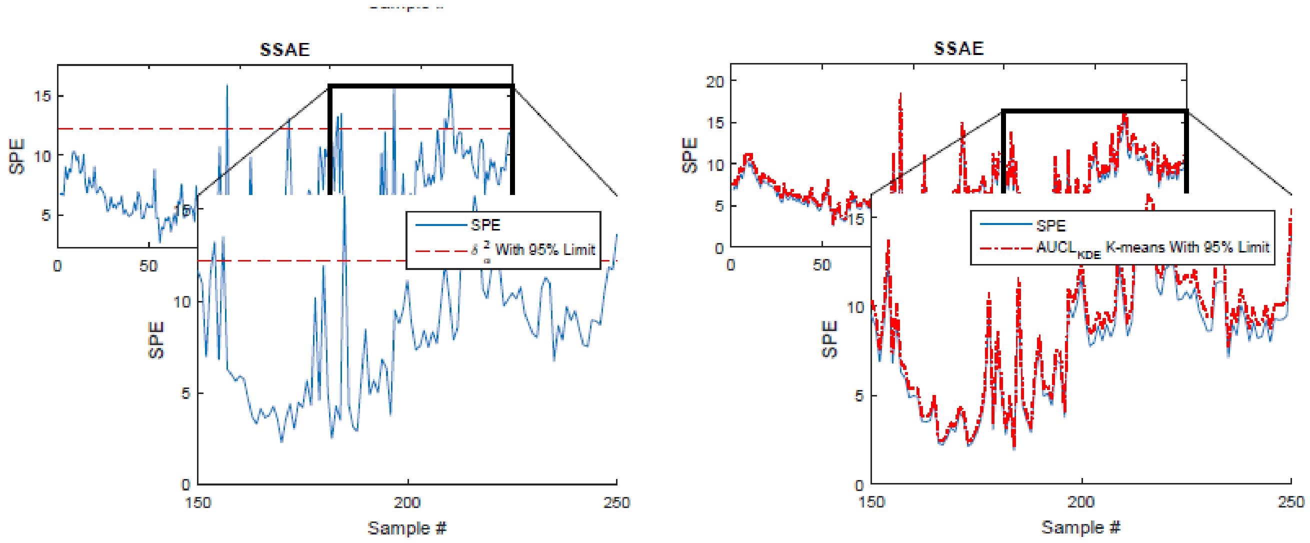


Fig.8a: Statistical threshold

Fig.8b: AUCL threshold using K-means clustering

Fig.8. SPE: data in normal state

We inject a fault in one of the sensors and we observe the evolution of the SPE where is shown in fig.9.

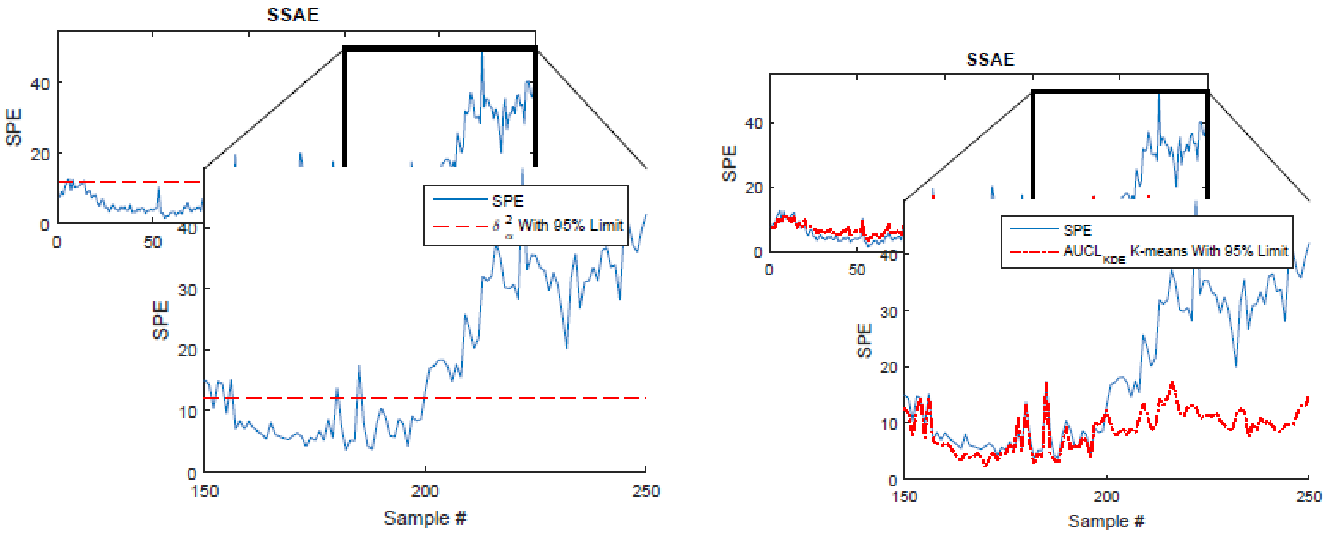


Fig.9a: Statistical threshold

Fig.9b: AUCL threshold using K-means clustering

Fig.9. SPE: data in faulty state

The SSAE model is able to detect and isolate the fault, its detection accuracy was from the 200 sample and the location of the defective sensor was the 8 sensor which is the O2 treated water (TW) as shown in fig.10 using normalized contribution plots and SVI method in fig.11, where this method has the advantage of reconstructing the erroneous sensor measurements.

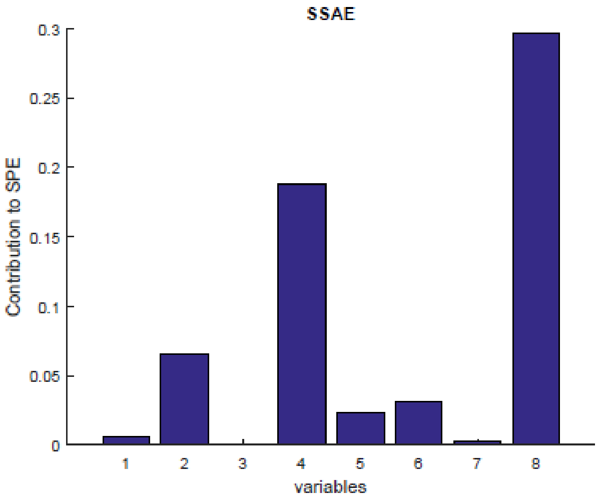


Fig. 10: Fault identification using normalized contribution Plots

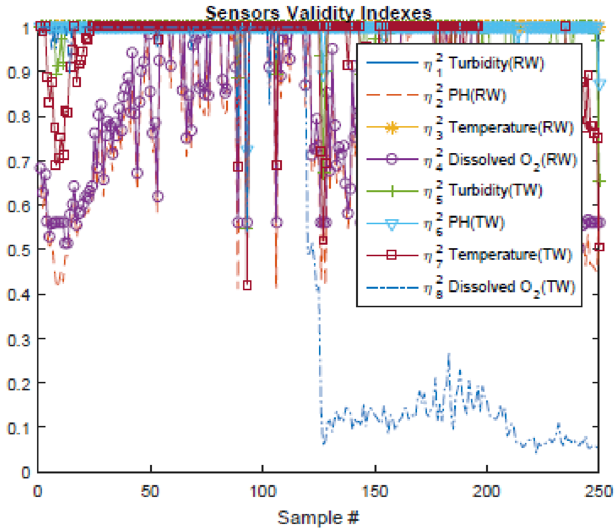


Fig. 11: SVI's Raw Water (RW) and Treated Water (TW) Sensors

*5.3 Scalability for Complex and High-Dimensional Systems*

As industrial systems grow in complexity, the volume and dimensionality of data generated by sensors also increase. The scalability of the proposed SSAE framework is a critical factor in its broader adoption for process monitoring.

1. Challenges in Scaling SSAE

- High Dimensionality: Monitoring systems with hundreds or thousands of sensors increases computational demands during both training and inference stages. The model may struggle to efficiently process such high-dimensional inputs without performance degradation.
- Real-Time Monitoring: Larger systems require real-time anomaly detection and fault isolation, necessitating a balance between computational efficiency and accuracy.
- Data Redundancy and Sparsity: High-dimensional datasets often include redundant or sparsely correlated features, which can dilute the model's ability to focus on critical variables.

2. Proposed Modifications and Optimizations

- Dimensionality Reduction:

Before feeding the data into the SSAE, we apply advanced feature extraction or dimensionality reduction techniques, such as t-SNE or autoencoders with bottleneck layers. This step reduces the computational burden without sacrificing critical information.

- Hierarchical Model Architecture:

For extremely large systems, we design a hierarchical monitoring structure. Divide the sensors into subsystems, train separate SSAEs for each subsystem, and aggregate their outputs through a meta-monitoring layer. This approach allows for localized anomaly detection and fault isolation while maintaining global oversight.

- Distributed and Parallel Processing:

We implement distributed training and inference using frameworks like TensorFlow Distributed or PyTorch Distributed. This ensures that the increased computational load is spread across multiple processing units, improving speed and efficiency.

- Optimized Training Techniques:

- ✓ We use batch normalization to stabilize and accelerate training on high-dimensional data.
- ✓ Implement transfer learning to fine-tune pre-trained SSAE models for specific subsystems, reducing the time and resources required for retraining.
- ✓ Employ adaptive learning rates and optimization techniques, such as Adam or RMSProp, to handle variability in training data more effectively.
- Sparse Regularization:

Enhance sparsity constraints in the hidden layers to focus the model's learning capacity on the most significant variables, thereby improving interpretability and reducing overfitting.

### 3. Experimental Validation

Future work could involve testing the scalability of the framework using large datasets from diverse industries, such as power grids or smart manufacturing plants. Metrics like training time, inference speed, and fault detection accuracy should be evaluated to validate the effectiveness of these optimizations.

### 4. Practical Implications

With these modifications, SSAE becomes a viable tool for real-time monitoring in complex industrial systems. It ensures scalability without compromising performance, making it suitable for applications such as:

- Monitoring nationwide utility grids.
- Supervising interconnected production lines in smart factories.
- Managing large-scale environmental monitoring systems.

By adopting these strategies, the proposed SSAE framework can evolve into a robust solution capable of handling the challenges posed by modern industrial systems.

## 6. Conclusion

This paper presents a novel framework using Stacked Sparse Autoencoders (SSAE) for anomaly detection and fault isolation in industrial processes. The approach addresses challenges related to nonlinear and dynamic systems, demonstrating its potential to enhance the reliability and efficiency of autonomous systems. Through the use of advanced techniques such as the Squared Prediction Error (SPE) index and Sensor Validity Index (SVI), the framework offers robust, adaptive monitoring solutions for sensor fault management.

Validated through experiments with synthetic and real-world data from a water treatment plant, the SSAE framework outperforms traditional methods like PCA, delivering higher accuracy in fault detection and isolation. The results underscore SSAE's capability to handle complex industrial environments and support real-time decision-making.

Looking ahead, the work will focus on scaling the model for larger, high-dimensional systems and improving computational efficiency, making it a key tool for the future of intelligent, autonomous industrial monitoring systems.

## References

[1] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2065, p. 20150202, mar 2016.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] G. E . Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," science, vol. 313, no. 5786, pp. 504–507,2006.

[3] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi,"Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," IEEE Transactions on Medical Imaging, vol. 35, no. 1, pp. 119–130, jan 2016..

[4] D. White, B. Goodlin, A. Gower, D. Boning, H. Chen, H. Sawin, and

T. Dalton, "Low open-area endpoint detection using a PCA-based t/sup statistic and q statistic on optical emission spectroscopy measurements,"IEEE Transactions on Semiconductor Manufacturing, vol. 13, no. 2, pp. 193–207, may 2000.

[5] P. Nomikos and J. F. MacGregor, "Multivariate spc charts for monitoring batch processes," Technometrics, vol. 37, no. 1, pp. 41–59, 1995.

[6] *Wafa Bougheloum, Mounir Bekaik, Sofiane Gherbi*, Multimode system condition monitoring using sparsity reconstruction for quality control, *International Journal of Electrical and Computer Engineering(IJECE)* p-ISSN 2088-8708, e-ISSN 2722-2578, 2022

[7] R. Dunia, S. J. Qin, T. F. Edgar, and T. J. McAvoy, "Use of principal

component analysis for sensor fault identification," Computers & Chemical Engineering, vol. 20, pp. S713–S718, jan 1996.

[8] K. Bouzenad, M. Ramdani, N. Zermi, and K. Mendaci, "Use of NLPCA for sensors fault detection and localization applied at WTP," in 2013 World Congress on Computer and Information Technology (WCCIT). IEEE, jun 2013.

[9] X. Zhang and Y. Li, "Multiway principal polynomial analysis for semiconductor manufacturing process fault detection," Chemometrics and Intelligent Laboratory Systems, vol. 181, pp. 29–35, oct 2018.

[10] K. Mendaci, M. Ramdani, and T. Benzaraa, "Nonlinear multivariate statistical process monitoring of a water treatment plant," in 2013

# New mapping method based on Bat algorithm for Embedded Systems[⋆]

Farid Boumaza[1,2,*,†], Atmane Hadji[3,†], Abdelkader Aroui[4,†] and Abou El hassan Benyamina[2,5,†]

[1] *Computer Science Department, University of Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj 34030, Algeria*

[2] *(LAPECI) Laboratory of Parallel, Embedded architectures and Intensive Computing, University of Oran1, Oran 31000, Algeria*

[3] *LISI Laboratory, Computer Science Department, University Center A. Boussouf Mila, 43000 Mila, Algeria*

[4] *Center for Space Techniques, Palestine Avenue, 31200 Arzew, Oran, Algeria*

[5] *Computer Science Department, University of Oran1, Oran 31000, Algeria*

## Abstract

The increasing complexity of Multi-Processor Systems-on-Chip (MPSoCs) platforms necessitates efficient task mapping strategies to optimize energy consumption and minimize latency. In this study, we propose a novel Binary Multi-Objwctive Bat Algorithm (BMOBA)-based mapping approach tailored for MPSoC platforms. The BMOBA leverages the echolocation-inspired behavior of bats, adapted for binary search spaces, to allocate tasks across processing elements while optimizing conflicting objectives. Our method models energy consumption and latency as the primary fitness functions and incorporates constraints such as load balancing, communication bandwidth, and memory availability. Extensive simulations demonstrate that the BMOBA outperforms traditional optimization techniques, such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), achieving significant reductions in both energy consumption and latency. This work highlights the potential of biologically inspired metaheuristics for addressing multi-objective optimization challenges in embedded systems and sets a foundation for future research in real-time applications on MPSoC platforms.

## Keywords

Binary Bat Algorithm, Multi-Processor Systems-on-Chip, Mapping, Energy Optimization

## 1. Introduction

As the demand for high-performance embedded systems continues to grow, Multi-Processor Systems-on-Chip (MPSoCs) have emerged as a cornerstone of modern computing platforms [1]. These architectures integrate multiple heterogeneous processing elements (PEs) onto a single chip, interconnected through advanced Network-on-Chip (NoC) communication systems. MPSoCs [2] are widely deployed in applications ranging from multimedia processing to real-time systems, where optimizing energy consumption and ensuring low latency are critical design objectives. However, achieving an efficient mapping of application tasks onto such architectures remains a challenging and computationally intensive problem [3, 4].

The mapping process involves assigning computational tasks to processing elements and scheduling communication across NoC links. This process must address multiple, often conflicting, objectives such as minimizing energy consumption, reducing communication delays, balancing load across PEs, and adhering to memory and bandwidth constraints [5]. As the problem is NP-hard, heuristic and metaheuristic approaches are increasingly employed to achieve near-optimal solutions within practical timeframes.

*Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY*

[⋆] You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

[*] Corresponding author.

[†] These authors contributed equally.

f.boumaaza@univ-bba.dz (F. Boumaza); a.hadji@centre-univ-mila.dz (A. Hadji); aroui_kader@yahoo.fr (A. Aroui); benyanabou@gmail.com (A. E. h. Benyamina)

0000-0002-9785-420X (F. Boumaza); 0000-0001-6706-6360 (A. Hadji); 0000-0003-4778-0123 (A. E. h. Benyamina)

In recent years, bio-inspired optimization algorithms [6] have garnered significant attention for solving complex, multi-objective problems in engineering domains. Among these, the Bat Algorithm (BA) [7], inspired by the echolocation behavior of microbats, has shown promise due to its adaptability, simplicity, and efficiency in navigating high-dimensional search spaces. While the traditional BA has primarily been applied to continuous optimization problems, its binary variant (Binary Bat Algorithm, BBA) [8] extends its applicability to discrete problems such as task mapping in MPSoC platforms. By modeling task allocation as a binary decision problem, the BBA offers a novel and effective solution for addressing energy consumption and latency optimization simultaneously.

This paper presents a Binary Multi-Objective Bat Algorithm (BMOBA) based mapping methodology for optimizing energy and latency in MPSoC platforms. The proposed approach formulates task mapping as a multi-objective optimization problem, where energy and latency are minimized subject to system constraints such as load balancing, bandwidth, and memory availability. The BBA is tailored to multi-Objective [9, 10, 11], and the binary nature of task assignment, enabling efficient exploration of the solution space and rapid convergence to near-optimal mappings.

To validate the effectiveness of the proposed approach, we conducted extensive experiments using standard benchmark applications and compared the results against established optimization techniques such as Non-Dominated Sorting Genetic Algorithms (NSGAII) [12] and Multi-Objective Particle Swarm Optimization (MOPSO) [13]. The experimental results demonstrate that the BMOBA achieves superior performance in terms of energy savings and latency reduction, making it a compelling choice for task mapping in MPSoC platforms.

The rest of the paper is organized as follows: Section 3 discusses the definitions and the necessary mathematical formulations for problem mapping onto the MPSoCs architecture. In Section 4, the proposed mapping strategies are presented. Experimental results are provided in Section 5, and the paper concludes with Section 6.

## 2. Related work

The mapping of tasks in MPSoC (Multiprocessor System-on-Chip) platforms is a central problem for optimizing resource utilization and ensuring high performance. This problem involves determining the optimal placement of tasks and communications on the platform's hardware resources, considering optimization criteria such as reducing energy consumption, minimizing total execution time, and optimizing channel occupancy. The goal is to efficiently distribute tasks among processors and manage data exchanges while respecting the platform's specific constraints.

The mapping techniques are particularly popular and have been the subject of extensive research in the literature, as they provide a well-defined and deterministic solution for task placement from the design phase. These techniques are effective in environments where workload and communication patterns are well-defined in advance. Significant search work has been done in this area, notably by Zhang(2002) [14], Shin(2004) [15], Ghosh(2009) [16], Vardi(2009) [15], Zadrija(2010) [17], Wang(2010) [18], Wang(2011) [19], Sahu(2013) [20], and Quan(2015) [21] which focus on task scheduling in MPSoC architectures using classical optimization methods.

Among the most commonly used optimization methods, metaheuristics have gained popularity due to their ability to handle complex problems and provide quality solutions. Genetic algorithms have been widely applied to solve the task mapping problem, as demonstrated by Wu(2003) [22], Lei(2003) [23], and , Boumaza(2012) [24]. These approaches exploit natural evolution mechanisms to explore the solution space for possible placements. Other metaheuristic techniques such as tabu search [25, 26] and simulated annealing [27, 28] have also been successfully used. These algorithms are designed to explore the solution space effectively, avoiding local minima and ensuring better exploration of the search space.

Recently, research has shifted toward multi-objective optimization techniques, particularly for solving complex problems where multiple performance criteria must be simultaneously optimized, such as energy consumption, execution time, and channel occupancy. The introduction of hybrid metaheuristic

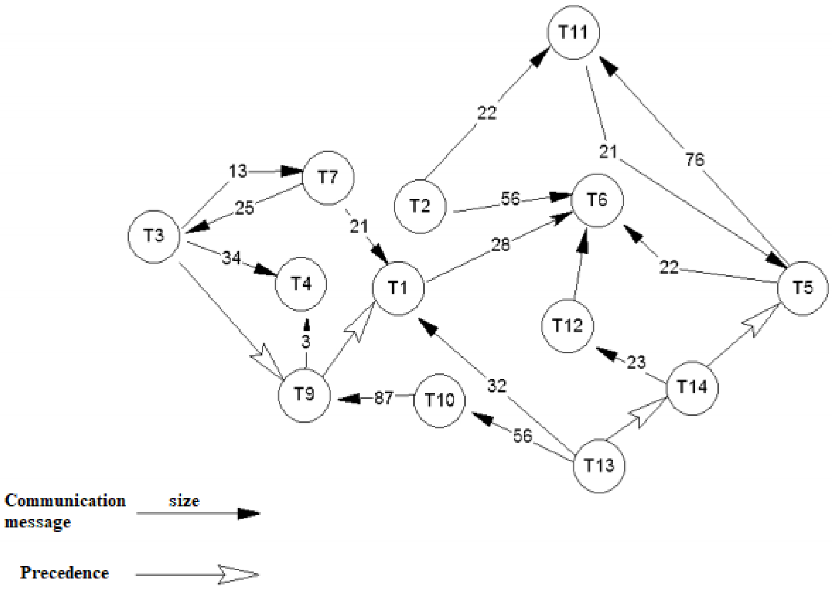


**Figure 1:** Application graph.

methods, such as the Bat Algorithm, addresses the limitations of traditional approaches by offering better exploration of the search space while maintaining rapid convergence toward optimal solutions.

## 3. Definition and formulation

Communications between the tasks of our application and the components of our target architecture is represented by two directed graphs.

**Definition 1**
The application graph, also known as the Task Graph (TG), is a directed graph $G(T, E)$, where each vertex $t_i \in T$ represents a module or task within the application. Each directed edge $(t_i, t_j)$, denoted as $e_{ij} \in E$, signifies a communication link between tasks $t_i$ and $t_j$. The weight of the edge $e_{ij}$, represented by $Q_{ij}$, indicates the volume of data transferred between $t_i$ and $t_j$, reflecting communication demand and aiding in optimizing resource allocation (see Fig. 1).

**Definition 2**
The architecture graph, denoted as $AG$ (Architecture Graph), is a directed graph $P(S, F)$ where each vertex $s_i \in S$ represents a node within the topology. The directed edge $(s_i, s_j)$, denoted as $f_{ij} \in F$, signifies a physical link that directly connects two elements, $s_i$ and $s_j$, within the architecture. The weight of the edge $f_{ij}$, represented by $bw_{ij}$, encapsulates critical characteristics of the physical link, including bandwidth, latency, and energy consumption. This comprehensive representation allows for effective analysis and optimization of communication pathways in Multi-Procesur-on-Chip (MPSoC) architectures (see Fig. 2).

**Definition 3**
The mapping of the application graph $G(T, E)$ onto the architecture graph $P(S, F)$ is defined by the mapping function:

$$\text{map} : T \to S \quad \text{such that} \quad \text{map}(t_i) = s_j \quad \forall\, t_i \in T\,, \exists\, s_j \in S. \tag{1}$$

This mapping is valid under the condition that the number of tasks $|T|$ is greater than or equal to the number of processing elements $|S|$ (see Fig. 3) [29].

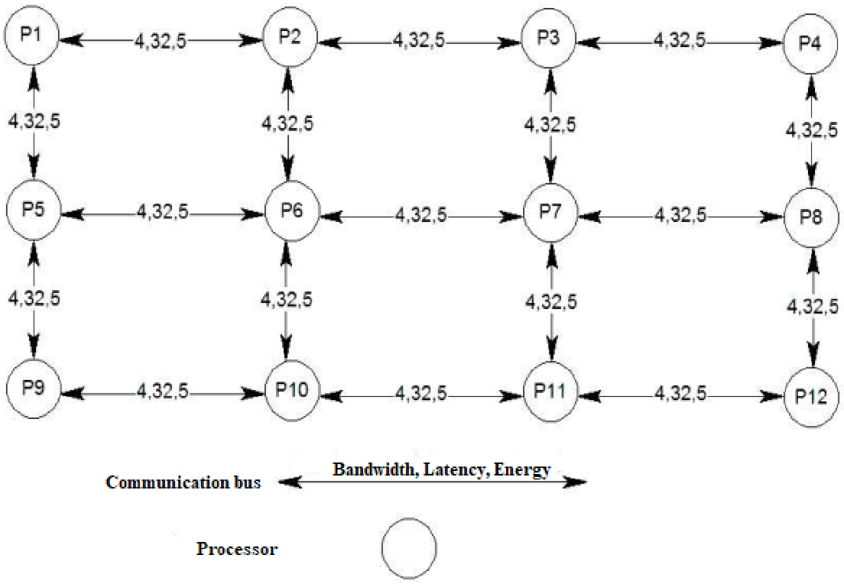


**Figure 2:** Architecture graph.

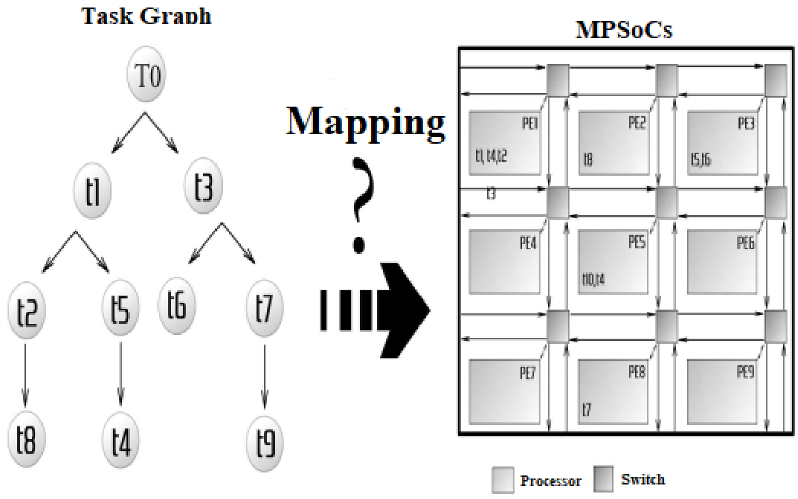


**Figure 3:** Mapping graph task on MPSoCs architecture.

The mapping process is crucial in optimizing resource allocation within MPSoCs, as it determines how application tasks are distributed across the available processing elements. An effective mapping strategy not only facilitates parallel execution of tasks but also minimizes communication overhead, ensuring that system performance and energy efficiency are maximized [3]. This approach is especially important in heterogeneous architectures, where diverse processing capabilities must be leveraged to meet the demands of complex applications.

Our application is defined as a set of tasks $T = \{t_1, t_2, \ldots, t_n\}$, while the target architecture is represented as a set of processors $P = \{p_1, p_2, \ldots, p_m\}$. It is important to note that each processor can operate in multiple modes, denoted as $m_1, m_2, m_3$. This capability of processors to function in various modes introduces greater diversity in optimization strategies for task placement.

By enabling processors to adapt their operational modes based on the specific requirements of tasks, we can achieve more efficient resource utilization and improved performance. This multi-mode functionality allows for fine-tuning of processing capabilities, enabling better handling of both regular and irregular task types. As a result, it facilitates enhanced flexibility in mapping strategies, leading to optimized execution times and reduced energy consumption within the multiprocessor system-on-chip (MPSoC) framework [2].

### 3.1. Execution Time and Communication Duration

The execution time $D$of a task is defined as follows:

$$D_i = \frac{\text{Taille}(t_i^p)}{f_{mp}} \tag{2}$$

where: - $D_i$ is the execution time of task $i$ on processor $p$. - Taille$(t_i^p)$ denotes the size of task $i$ for processor $p$. - $f_{mp}$ is the frequency of processor $p$ operating in mode $m$.

The overall execution time for the application is given by:

$$D = \max(D_i) \quad \text{for} \quad i = 1, \ldots, \text{number of tasks} \tag{3}$$

where $D$ is the maximum execution time among all tasks $D_i$, taking into account their dependencies.

The duration of communication $D_{com}^{ij}$ between tasks $i$ and $j$ is calculated as:

$$D_{com}^{ij} = \frac{Q_{ij}}{\text{Min } B_{P|p_k,p_l|}} \times \sum \text{latency}(p_k, p_l) \tag{4}$$

where: - $D_{com}^{ij}$ is the communication duration between tasks $i$ and $j$. - $Q_{ij}$ represents the data volume that needs to be communicated between tasks $i$ and $j$. - Min $B_{P|p_k,p_l|}$ denotes the minimum bandwidth of the path connecting processors $p_k$ and $p_l$. - The summation $\sum$ latency$(p_k, p_l)$ accounts for the cumulative latency across the communication path.

This approach ensures that both execution and communication times are adequately considered for optimizing task mapping in MPSoCs, thereby improving overall system performance and efficiency.

### 3.2. Energy Consumption

The energy consumption $E$ in a multiprocessor system-on-chip (MPSoC) can be categorized into execution energy and communication energy.

#### 3.2.1. Execution Energy

The energy consumed during the execution of a task $i$ on processor $p$ in mode $m$ is defined as:

$$E_{\text{exec}}^i = \text{Size}_{ip} \times e_{mp} \tag{5}$$

where:

- $E_{\text{exec}}^i$ is the execution energy of task $i$.
- Size$_{ip}$ represents the number of cycles required for task $i$ to execute on processor $p$ in mode $m$.
- $e_{mp}$ denotes the energy consumption per cycle for processor $p$ in mode $m$.

#### 3.2.2. Communication Energy

The energy consumed due to communication between tasks $i$ and $j$ assigned to processors $p$ and $q$ is given by:

$$E_{\text{com}}^{ijpq} = \sum_{i=1}^{n} Q_{ij} \times e_{p_l,p_k} \tag{6}$$

where:

- $E_{\text{com}}^{ijpq}$ is the communication energy between tasks $i$ and $j$.

- $Q_{ij}$ represents the volume of data exchanged between tasks $i$ and $j$.
- $e_{p_l,p_k}$ is the energy cost associated with the communication link between processors $p_l$ and $p_k$.

Where:

- $E_{\text{total}}$ is the overall energy consumption for all tasks in the application.
- The summation encompasses both the execution energy and communication energy across all tasks.

#### 3.2.3. Total Energy Consumption

The total energy consumption for executing all tasks and their communications within the system can be expressed as:

$$E_{\text{total}} = \sum_{i=1}^{\text{task}_n} (E_{\text{exec}^i} + E_{\text{com}}^{ijpq}) \tag{7}$$

### 3.3. Task Placement Indicator

An auxiliary variable $X_m^{ip}$ is used to indicate the placement of tasks on processors, defined as follows:

$$X_m^{ip} = \begin{cases} 1 & \text{if task } i \text{ is placed on processor } p \text{ and operates in mode } m \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

This formulation allows for a comprehensive analysis of energy consumption during task execution and inter-task communication, facilitating the optimization of task mapping strategies in MPSoCs. By minimizing both execution and communication energy, the overall efficiency and sustainability of the system can be significantly enhanced.

## 4. Proposed Resolution Method

The proposed approach addresses the Assignment and Scheduling (AS) problem, which involves allocating application tasks and scheduling their associated communications onto the resources of a target architecture. The primary objective is to meet specified performance metrics while optimizing resource utilization.

Our methodology considers multiple objectives, including minimizing energy consumption and minimzing the time (latency). These objectives are critical for mobile embedded systems, where energy efficiency impacts battery longevity [30], and performance efficiency directly influences task execution speed. However, these goals often conflict—for instance, operating components in energy-saving modes can extend battery life but increase task completion times.

To address these conflicting objectives, we employ a multi-objective optimization strategy that balances energy consumption and performance. Specifically, we adapte the Multi-Objective Bat Algorithm (MOBA) technique to tackle the AS problem effectively. This approach achieves a trade-off between energy efficiency and execution speed, ensuring an optimized mapping of tasks in MPSoC platforms.

By integrating advanced optimization techniques, our approach offers a robust and scalable solution for the dynamic requirements of mobile embedded systems. It simultaneously addresses critical constraints, such as energy efficiency and task execution time, thereby enhancing the overall performance and sustainability of MPSoC architectures.

### 4.1. Binary Bat Algorithm (BBA)

The Binary Bat Algorithm (BBA) [8] is an adaptation of the Bat Algorithm (BA) [7] designed specifically to handle discrete optimization problems [31] within binary search spaces. Originally, the BA was developed to solve continuous optimization challenges [32] by mimicking the echolocation behavior of bats. However, many practical optimization problems, such as task mapping in MPSoCs, require solutions in binary spaces (e.g., assignment of tasks or resources, represented as 0s and 1s). The BBA introduces novel mechanisms to bridge this gap by redefining the core concepts of position and velocity updates in binary domains.

### Principles of the Binary Bat Algorithm

The BBA maintains the fundamental principles of the original BA while modifying its mathematical model to operate in binary search spaces. Its main characteristics include:

- **Position Representation:** Each bat represents a potential solution as a binary vector, where each bit can take the value of 0 or 1.
- **Velocity and Position Update:** In continuous BA, positions are updated by adding velocity vectors. In BBA, the velocity represents the probability of a position bit flipping ($0 \leftrightarrow 1$), which is governed by a transfer function.
- **Transfer Function:** The velocity is mapped to a probability using a transfer function, which determines the likelihood of flipping a bit. Common transfer functions include:
  - *Sigmoid Function:* Produces smooth probabilities for bit flipping.
  - *V-shaped Transfer Function:* Enhances exploration by increasing the flipping probability for larger velocities, ensuring better traversal of the search space.

### 4.1.1. Mathematical Formulation of Bat Algorithm

The core mechanics of the Bat Algorithm are based on the echolocation principles of bats, modeled mathematically as follows:

1. **Velocity Update:**
$$v_i^{t+1} = v_i^t + (x_i^t - x_{\text{best}}) f_i, \tag{9}$$
where $v_i^t$ is the velocity of bat $i$ at iteration $t$, $x_i^t$ is the position of bat $i$, $x_{\text{best}}$ is the global best solution, and $f_i$ is the frequency associated with bat $i$.
2. **Frequency Update:**
$$f_i = f_{\text{min}} + (f_{\text{max}} - f_{\text{min}}) \cdot \beta, \tag{10}$$
where $\beta$ is a random number drawn from a uniform distribution in $[0, 1]$, and $f_{\text{min}}$ and $f_{\text{max}}$ define the range of frequencies.
3. **Position Update:**
$$x_i^{t+1} = x_i^t + v_i^{t+1}. \tag{11}$$
In binary spaces, this update is performed probabilistically, as described in the next subsection.
4. **Loudness and Pulse Rate Update:**
$$A_i^{t+1} = \alpha \cdot A_i^t, \tag{12}$$
$$r_i^{t+1} = r_i^0 \cdot (1 - \exp(-\gamma t)), \tag{13}$$
where $\alpha$ and $\gamma$ are constants controlling the decay of loudness $A$ and the increase of pulse rate $r$, respectively.

### 4.1.2. Binary Search Space Adaptation

In BBA, the continuous position update formula is replaced with a binary mechanism:

1. The velocity $v_i^{t+1}$ is mapped to a probability using a transfer function $S(v)$:
$$S(v_i) = \frac{1}{1 + \exp(-v_i)}, \tag{14}$$
for the sigmoid function, or:
$$S(v_i) = \frac{2}{\pi} \arctan\left(\frac{\pi}{2} v_i\right), \tag{15}$$
for the V-shaped function.

2. The position $x_i$ is updated based on the probability $S(v_i)$:

$$x_i^{t+1} = \begin{cases} 1 - x_i^t, & \text{if rand} < S(v_i^{t+1}), \\ x_i^t, & \text{otherwise.} \end{cases} \tag{16}$$

#### 4.1.3. Algorithmic Steps

The BBA proceeds through the following steps:

1. **Initialization:** A population of bats is initialized with binary positions (solution vectors) and velocities.
2. **Frequency Adjustment:** Each bat's frequency is adjusted to guide exploration in the solution space, encouraging diversity.
3. **Velocity Update:** Velocities are updated based on the bat's current position, the best global solution, and a random frequency. These velocities are then mapped to probabilities using the transfer function.
4. **Position Update:** Bits in the position vector are flipped with a probability determined by the mapped velocity.
5. **Fitness Evaluation:** The quality of each bat's solution is evaluated using a fitness function defined by the optimization objectives (e.g., minimizing energy consumption and latency in MPSoCs).
6. **Exploration and Exploitation:** Loudness and pulse emission rate parameters are dynamically adjusted to balance global exploration and local exploitation.
7. **Convergence:** The process repeats until a termination condition, such as a maximum number of iterations or acceptable fitness level, is met.

#### 4.1.4. Adaptation for MPSoC Task Mapping

In this work, the BBA is employed to solve the task mapping problem on MPSoC platforms. The binary nature of the task assignment process aligns perfectly with the capabilities of BBA, enabling efficient optimization of conflicting objectives such as energy consumption and execution time.

### 4.2. Multi-Objective Bat Algorithm (MOBA)

The Multi-Objective Bat Algorithm (MOBA) [33] extends the Bat Algorithm (BA) to handle multi-objective optimization problems. Traditional BA focuses on single-objective optimization, but many real-world problems, such as task mapping in MPSoCs, require balancing conflicting objectives, such as minimizing energy consumption while reducing execution latency. MOBA incorporates strategies to explore trade-offs among multiple objectives, providing a Pareto-optimal set of solutions.

#### 4.2.1. Principles of MOBA

MOBA retains the core mechanics of BA while introducing mechanisms for multi-objective optimization:

- **Pareto Dominance:** Solutions are evaluated based on Pareto dominance, ensuring that non-dominated solutions are retained across iterations (See fig 5).
- **Archiving Mechanism:** An external archive is used to store the best non-dominated solutions found during the search process. This archive represents the Pareto front (See fig 4).
- **Crowding Distance:** To maintain diversity among solutions in the Pareto front, MOBA uses a crowding distance metric to prioritize solutions in less crowded regions.

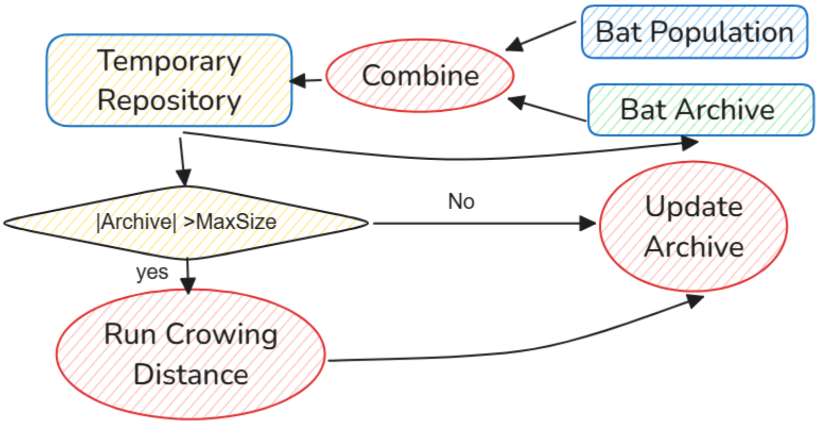


**Figure 4:** The management of the archive population.

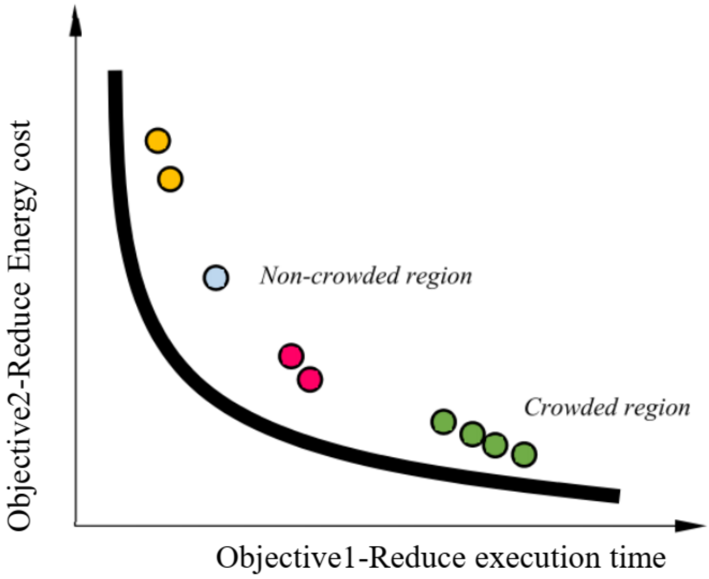


**Figure 5:** Managing solution diversity.

### 4.2.2. Algorithmic Steps of MOBA

The general process of MOBA can be summarized as follows:

1. **Initialization:** Initialize a population of bats with random positions and velocities, as well as an empty Pareto archive.
2. **Fitness Evaluation:** Evaluate each solution based on multiple objectives. Identify non-dominated solutions and update the Pareto archive.
3. **Velocity and Position Updates:** Update velocities and positions of bats using standard BA equations. In MOBA, solutions in the Pareto archive are used as references for guiding the search.
4. **Archiving and Diversity Management:** Add newly identified non-dominated solutions to the archive. Use crowding distance to manage archive size and maintain solution diversity.
5. **Convergence Check:** Repeat the process until a stopping criterion, such as a maximum number of iterations or convergence of the Pareto front, is met.

The placement and scheduling phase plays a pivotal role in our design flow, as it significantly impacts the implementation of an application on a specialized architecture. This phase processes the following inputs:

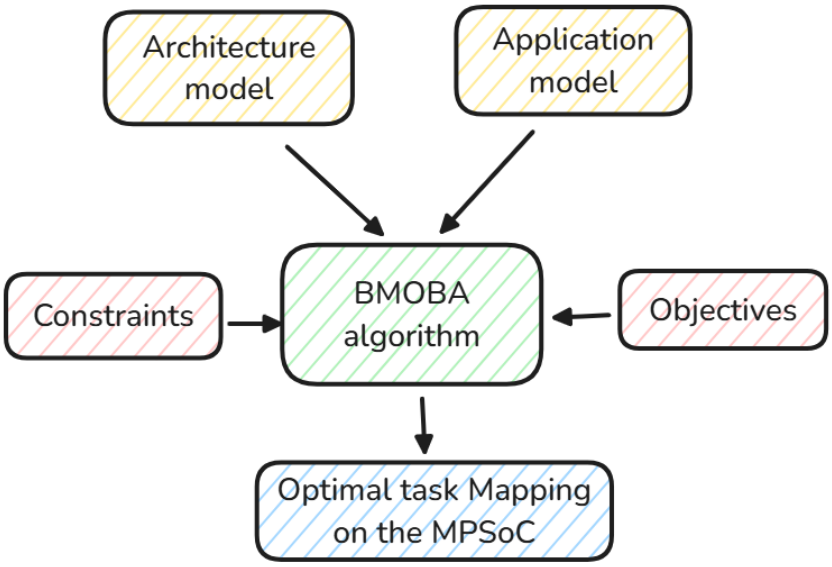


**Figure 6:** Global description of our solution.

- **Application Model**: A comprehensive representation of the application, detailing its structure, task dependencies, and communication requirements.
- **Target Architecture Model**: A representation of the hardware platform, describing the available resources, their configurations, and interconnections.
- **Performance and Energy Constraints**: Defined requirements that the implementation must satisfy, including execution time limits and energy consumption thresholds.
- **Objective Functions**: Optimization criteria such as minimizing latency, reducing energy usage, or achieving load balancing.

The outcome of this phase is an optimized mapping of tasks and communication flows to the physical resources of the architecture, along with a detailed scheduling strategy that ensures compliance with performance and energy constraints. A comprehensive depiction of our approach is illustrated in the the figure(6).

## 4.3. Convergence and Exploration-Exploitation Trade-offs

The Binary Multi-Objective Bat Algorithm (BMOBA) is designed to effectively balance exploration and exploitation to solve complex multi-objective optimization problems, such as task mapping in MPSoC systems. A key feature of BMOBA is its ability to navigate the search space by adjusting parameters like frequency, loudness, and pulse emission rate, which govern the behavior of the bats. This allows the algorithm to dynamically switch between exploring new areas of the solution space and exploiting promising solutions.

### 4.3.1. Convergence Analysis

BMOBA converges toward an optimal solution by gradually reducing the loudness and increasing the pulse rate over time. Early iterations focus on exploration by allowing bats to search widely, while later iterations shift towards exploitation to refine solutions. The convergence speed and solution quality are influenced by the dynamic tuning of these parameters, ensuring that BMOBA avoids premature convergence to local optima while still achieving high-quality results.

### 4.3.2. Trade-off Between Exploration and Exploitation

One of the main advantages of BMOBA is its ability to manage the trade-off between exploration and exploitation. Exploration is achieved by high-frequency adjustments, which allow bats to explore distant regions of the solution space, while exploitation is facilitated by reducing frequency and increasing loudness to fine-tune the solutions. This adaptive mechanism ensures that BMOBA maintains diversity in the population during the early stages of the search and gradually converges to optimal solutions.

### 4.3.3. Robustness

The robustness of BMOBA lies in its adaptive parameter adjustment, which helps the algorithm respond effectively to the complexity of the optimization problem. By adjusting the exploration-exploitation balance during the search process, BMOBA can adapt to different problem landscapes, ensuring that it remains effective even in the presence of multi-modal and highly-constrained objective spaces.

The dynamic adjustment of parameters such as loudness and frequency ensures that the algorithm can adapt to the characteristics of the problem being solved. For instance, in problems where the solution space is large or complex, the algorithm can maintain a high level of exploration. In problems where more precise solutions are required, BMOBA gradually shifts towards **exploitation**, refining the solutions found in earlier iterations.

## 5. Experimentation and Results

Our approach was implemented in Matlab, and all experiments were conducted on a system equipped with an Intel(R) Core(TM) i5-7300HQ CPU running Windows 10. Utilizing the binary MOBA-based mapping solution, configured with the parameters specified in Table 1, we obtained the following results:

**Table 1**
The basic paprameters of BMOGWO

| Algorithm | Basic parameters | |
|---|---|---|
| BMOBA | - Number of tasks | 21 |
| | - Number of processors | 8 |
| | - Architecture type | Star topology |
| | - Latency | 1 unit |
| | - Population size | 20 |
| | - Archive size | 20 |
| | - Number of iterations | 30 |
| | - $A$ | 0.25 |
| | - $r$ | 0.5 |
| | - $\epsilon$ | [-1 1] |

### 5.1. Comparison of BMOBA, MOPSO, and NSGA-II

To evaluate the effectiveness of our proposed Binary Multi-Objective Bat Optimizer (BMOBA) for mapping applications onto MPSoCs, we conducted a comparative analysis against two widely recognized multi-objective optimization algorithms: Multi-Objective Particle Swarm Optimization (MOPSO) and Non-dominated Sorting Genetic Algorithm II (NSGA-II). These comparisons were performed on a benchmark set of medium-sized applications, with variations in the number of processors and task quantities.

The evaluation focused on two critical metrics: execution time and energy consumption. Each algorithm's performance was assessed under diverse MPSoC configurations to capture their adaptability and efficiency in addressing mapping challenges. Table 2 provides a detailed summary of the results, showcasing BMOBA's superior ability to simultaneously optimize execution time and energy consumption.

The findings underline BMOBA's versatility and effectiveness in discrete optimization tasks, demonstrating its capability to handle the complex requirements of MPSoC environments. This study highlights the comparative advantages of BMOBA, offering valuable insights into its adaptability and efficiency relative to established techniques like MOPSO and NSGA-II.

**Table 2**
Comparison of BMOBA, MOPSO, and NSGA-II on MPSoC Mapping

| Processors | Tasks | BMOBA | | MOPSO | | NSGA-II | |
|---|---|---|---|---|---|---|---|
| | | Exec Time (s) | Energy (J) | Exec Time (s) | Energy (J) | Exec Time (s) | Energy (J) |
| 4 | 10 | 0.8 | 5.3 | 1.2 | 5.6 | 1.1 | 5.4 |
| 6 | 15 | 1.6 | 6.9 | 1.7 | 7.1 | 1.6 | 7.0 |
| 8 | 20 | 2.4 | 8.8 | 2.5 | 8.9 | 2.4 | 8.7 |
| 10 | 25 | 3.0 | 10.5 | 3.3 | 10.8 | 3.0 | 10.6 |
| 12 | 30 | 3.9 | 12.3 | 4.0 | 12.6 | 3.8 | 12.4 |
| 14 | 35 | 4.7 | 14.7 | 4.8 | 15.0 | 4.5 | 14.8 |
| 16 | 40 | 5.5 | 16.9 | 5.7 | 17.0 | 5.5 | 16.8 |
| 18 | 45 | 6.5 | 18.8 | 6.5 | 18.7 | 6.2 | 18.5 |
| 20 | 50 | 6.9 | 20.1 | 7.3 | 20.5 | 7.0 | 20.3 |

The results clearly demonstrate that BMOBA consistently outperforms MOPSO and NSGA-II, achieving lower execution times and reduced energy consumption, especially in scenarios involving higher task loads and larger processor counts. These findings highlight BMOBA's effectiveness as a robust solution for application mapping in MPSoC environments, particularly in cases where discrete optimization and energy efficiency are of paramount importance.

## 6. Conclusion

This paper presented a novel approach utilizing the multi-objective variant of the Bat Algorithm (BMOBA) to address the complex problem of mapping hierarchical real-time applications onto hierarchical MPSoC architectures. By incorporating a binary encoding scheme, our method effectively optimized two critical factors in real-time embedded systems: execution time and energy consumption. The experimental results, benchmarked against two widely used metaheuristic algorithms, demonstrated that our proposed solution consistently outperformed these alternatives in both execution time and energy efficiency. These results highlight the strength and adaptability of our approach for optimizing task mapping in MPSoC environments.

However, there are some limitations to the current approach that should be addressed in future work. One key limitation is the scalability of the proposed method. While our results are promising for small to medium-sized problems, the algorithm's performance and efficiency in handling larger and more complex MPSoC systems with numerous tasks and processors need further investigation. As the size of the system increases, the search space grows exponentially, potentially leading to longer computation times and reduced effectiveness in finding optimal solutions. Future work could focus on improving the scalability of the algorithm, possibly by introducing parallelization.

## References

[1] A. Schranzhofer, J.-J. Chen, L. Thiele, Dynamic power-aware mapping of applications onto heterogeneous mpsoc platforms, IEEE Transactions on Industrial Informatics 6 (2010) 692–707.

[2] W. Wolf, A. A. Jerraya, G. Martin, Multiprocessor system-on-chip (mpsoc) technology, IEEE transactions on computer-aided design of integrated circuits and systems 27 (2008) 1701–1713.

[3] A. Aroui, P. Boulet, K. Benhaoua, A. K. Singh, et al., Novel metric for load balance and congestion reducing in network on-chip, Scalable Computing: Practice and Experience 21 (2020) 309–321.

[4] T. Noergaard, Embedded systems architecture: a comprehensive guide for engineers and programmers, Newnes, 2012.

[5] Y. R. Muhsen, N. A. Husin, M. B. Zolkepli, N. Manshor, A. A. J. Al-Hchaimi, H. M. Ridha, Enhancing noc-based mpsoc performance: a predictive approach with ann and guaranteed convergence arithmetic optimization algorithm, IEEE Access (2023).

[6] A. C. Johnvictor, V. Durgamahanthi, R. M. Pariti Venkata, N. Jethi, Critical review of bio-inspired optimization techniques, Wiley Interdisciplinary Reviews: Computational Statistics 14 (2022) e1528.

[7] X.-S. Yang, A. Hossein Gandomi, Bat algorithm: a novel approach for global engineering optimization, Engineering computations 29 (2012) 464–483.

[8] S. Mirjalili, S. M. Mirjalili, X.-S. Yang, Binary bat algorithm, Neural Computing and Applications 25 (2014) 663–681.

[9] F. Boumaaza, A. E. H. Benyamina, Mapping multi objectifs d ‘application intensive sur architecture mpsoc (2012).

[10] F. Boumaza, A. E. H. Benyamina, D. Zouache, L. Abualigah, A. Alsayat, An improved harris hawks optimization algorithm based on bi-goal evolution and multi-leader selection strategy for multi-objective optimization., Ingénierie des Systèmes d'Information 28 (2023).

[11] N. Gunantara, A review of multi-objective optimization: Methods and its applications, Cogent Engineering 5 (2018) 1502242.

[12] K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii, in: Parallel Problem Solving from Nature PPSN VI: 6th International Conference Paris, France, September 18–20, 2000 Proceedings 6, Springer, 2000, pp. 849–858.

[13] Q. Lin, J. Li, Z. Du, J. Chen, Z. Ming, A novel multi-objective particle swarm optimization with multiple search strategies, European Journal of Operational Research 247 (2015) 732–744.

[14] Y. Zhang, X. S. Hu, D. Z. Chen, Task scheduling and voltage selection for energy minimization, in: Proceedings of the 39th annual Design Automation Conference, 2002, pp. 183–188.

[15] D. Shin, J. Kim, Power-aware communication optimization for networks-on-chips with voltage scalable links, in: Proceedings of the 2nd IEEE/ACM/IFIP international conference on Hardware/-software codesign and system synthesis, 2004, pp. 170–175.

[16] P. Ghosh, A. Sen, A. Hall, Energy efficient application mapping to noc processing elements operating at multiple voltage levels, in: 2009 3rd ACM/IEEE International Symposium on Networks-on-Chip, IEEE, 2009, pp. 80–85.

[17] V. Zadrija, V. Sruk, Mapping algorithms for mpsoc synthesis, in: The 33rd International Convention MIPRO, IEEE, 2010, pp. 624–629.

[18] X. Wang, M. Yang, Y. Jiang, P. Liu, Power-aware mapping for network-on-chip architectures under bandwidth and latency constraints, in: 2009 Fourth International Conference on Embedded and Multimedia Computing, IEEE, 2009, pp. 1–6.

[19] F. Wang, Y. Chen, C. Nicopoulos, X. Wu, Y. Xie, N. Vijaykrishnan, Variation-aware task and communication mapping for mpsoc architecture, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 30 (2011) 295–307.

[20] P. K. Sahu, S. Chattopadhyay, A survey on application mapping strategies for network-on-chip design, Journal of systems architecture 59 (2013) 60–76.

[21] W. Quan, A. D. Pimentel, A hybrid task mapping algorithm for heterogeneous mpsocs, ACM Transactions on Embedded Computing Systems (TECS) 14 (2015) 1–25.

[22] D. Wu, B. M. Al-Hashimi, P. Eles, Scheduling and mapping of conditional task graph for the synthesis of low power embedded systems, IEE Proceedings-Computers and Digital Techniques 150 (2003) 262–273.

[23] T. Lei, S. Kumar, Algorithms and tools for network on chip based system design, in: 16th Symposium on Integrated Circuits and Systems Design, 2003. SBCCI 2003. Proceedings., IEEE,

2003, pp. 163–168.

[24] F. Boumaaza, A. E. H. Benyamina, Mapping multi objectifs d ‘application intensive sur architecture mpsoc (2012).

[25] S. Manolache, P. Eles, Z. Peng, Fault and energy-aware communication mapping with guaranteed latency for applications implemented on noc, in: Proceedings of the 42nd annual Design Automation Conference, 2005, pp. 266–269.

[26] S. Murali, M. Coenen, A. Radulescu, K. Goossens, G. De Micheli, A methodology for mapping multiple use-cases onto networks on chips, in: Proceedings of the Design Automation & Test in Europe Conference, volume 1, IEEE, 2006, pp. 1–6.

[27] C. Marcon, A. Borin, A. Susin, L. Carro, F. Wagner, Time and energy efficient mapping of embedded applications onto nocs, in: Proceedings of the 2005 Asia and South Pacific Design Automation Conference, 2005, pp. 33–38.

[28] H. Orsila, T. Kangas, E. Salminen, T. D. Hämäläinen, M. Hännikäinen, Automated memory-aware application distribution for multi-processor system-on-chips, Journal of Systems Architecture 53 (2007) 795–815.

[29] K. Laredj, M. Belarbi, A. E. Benyamina, Metrics for real-time solutions design, in: Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2, Springer, 2019, pp. 411–425.

[30] A. Mehran, S. Saeidi, A. Khademzadeh, A. Afzali-Kusha, Spiral: A heuristic mapping algorithm for network on chip, IEICE Electronics Express 4 (2007) 478–484.

[31] J. Krause, J. Cordeiro, R. S. Parpinelli, H. S. Lopes, A survey of swarm algorithms applied to discrete optimization problems, in: Swarm intelligence and bio-inspired computation, Elsevier, 2013, pp. 169–191.

[32] Z.-H. Zhan, L. Shi, K. C. Tan, J. Zhang, A survey on evolutionary computation for complex continuous optimization, Artificial Intelligence Review 55 (2022) 59–110.

[33] L. L. Laudis, S. Shyam, C. Jemila, V. Suresh, Moba: multi objective bat algorithm for combinatorial optimization in vlsi, Procedia Computer Science 125 (2018) 840–846.

# Optimizing E-Commerce Product Recommendations Using Reinforcement Learning

Abderaouf Bahi[1, *, †], Ibtissem Gasmi[1, †], Sassi Bentrad[2, †], Ramzi Khantouchi[1, †]

[1] *Computer Science and Applied Mathematics Laboratory, Chadli Bendjedid El Tarf University, El Tarf 36000, Algeria*
[2] *LISCO Laboratory, National Higher School of Cyber Security (NSCS), Algers 16000, Algeria*

**Abstract**

In order to improve the user experience by offering personalized product recommendations, e-commerce platforms necessitate sophisticated recommendation systems. Reinforcement learning (RL) provides a more dynamic solution by adapting to real-time user interactions, in contrast to traditional recommendation systems that rely on demographic data and consumption history. This paper introduces a recommendation model that is based on reinforcement learning and is applied to the Amazon Product dataset. The model optimizes product suggestions by considering user preferences. We investigate the potential of RL to enhance the accuracy of recommendations in comparison to baseline methods, including content-based techniques and collaborative filtering. The model demonstrated the efficacy of RL in improving e-commerce product recommendations by achieving superior accuracy and recall. Our results indicate that reinforcement learning has the potential to considerably enhance the relevance of recommendations, thereby resulting in a superior user experience. By incorporating supplementary data sources and increasing algorithmic complexity, the proposed approach has the potential to be further enhanced.

**Keywords**

Reinforcement Learning, E-commerce, Recommendation Systems, Amazon Product Dataset.

## 1. Introduction

The way consumers shop has been transformed by e-commerce platforms, which provide a vast array of products that can be accessed with the press of a button. Nevertheless, the sheer volume of available items can be overwhelming for consumers, resulting in a challenge in identifying products that align with their preferences. In order to confront this obstacle, recommendation systems (RS) have emerged as indispensable instruments in the realm of e-commerce. These systems offer consumers customized recommendations that are informed by their browsing habits, purchase history, and preferences. The objective of these systems is to enhance the user experience, increase engagement, and, in the end, increase sales for businesses by promoting items that are in accordance with user interests.

Content-based filtering and collaborative filtering [1, 2] are the two primary techniques that traditional recommendation systems primarily rely on. Item attributes, including descriptions and categories, are analyzed by content-based filtering to suggest products that are similar to those with which a user has previously interacted. On the other hand, collaborative filtering capitalizes on the preferences of users with similar interests by suggesting items that they have

*ICASA'24: 2nd International Conference On Autonomous systems and their application, 12–15, 2024, El Taref University*

* Abderaouf Bahi.
† These authors contributed equally.
a.bahi@univ-eltarf.dz (A. Bahi); gasmi-ibtissem@univ-eltarf.dz (I. Gasmi); sassi_bentrad@hotmail.fr (S. Bentradd); r.khantouchi@univ-eltarf.dz (R. Khantouchi)
0009-0003-7116-5080 (A. Bahi); 0000-0002-8939-1727 (I. Gasmi); 0000-0002-7458-8121 (S. Bentrad); 0009-0009-5391-9022

Content-based filtering and collaborative filtering [1, 2] are the two primary techniques that traditional recommendation systems primarily rely on. Item attributes, including descriptions and categories, are analyzed by content-based filtering to suggest products that are similar to those with which a user has previously interacted. On the other hand, collaborative filtering capitalizes on the preferences of users with similar interests by suggesting items that they have enjoyed. While these methodologies have been effective in a variety of applications, they are subject to significant limitations. For instance, collaborative filtering frequently encounters the cold-start issue, which results in recommendations that are less precise due to the absence of sufficient interaction data for new users or items. In the same vein, content-based filtration may not be able to capture the complex patterns of user-item interactions, which could lead to repetitive suggestions and a decrease in the diversity of recommendations.

The necessity for more advanced, adaptive recommendation techniques has increased as e-commerce continues to expand. Reinforcement learning (RL), a subfield of machine learning [3, 4], has recently emerged as a promising approach to address some of the inherent limitations of traditional RS. RL models are distinguished from conventional approaches by their emphasis on real-time user interactions, which facilitate continuous learning. To optimize long-term rewards, RL models employ feedback mechanisms, which dynamically modify recommendations based on user behavior, rather than solely relying on historical data. This enables RL-based systems to enhance overall user satisfaction by providing the ability to predict future user preferences and increase immediate recommendation accuracy. In the context of e-commerce, where user preferences are constantly evolving and the product landscape is constantly changing, this capacity to learn and adapt renders RL particularly well-suited for recommendation systems.

This paper investigates the implementation of a reinforcement learning-based recommendation system on the Amazon Product dataset, a comprehensive dataset that includes a diverse array of products and user reviews. Our objective is to develop a model that can generate product recommendations that are more precise and diverse, and that are customized to the individual preferences of each user, through the use of RL. The strategic selection of the Amazon Product dataset is justified by its diversity and scope, which provide a realistic environment for testing the robustness of the RL approach. Furthermore, this dataset contains valuable user feedback in the form of ratings and evaluations, which can be utilized as a reward signal for the RL agent to assist in the development of more pertinent recommendations.

Our study builds upon previous research in both traditional and RL-based recommendation systems. Reinforcement learning has shown considerable promise in domains such as movie recommendations and online content personalization, where user feedback can be collected in real time to fine-tune the recommendation process. However, the application of RL in e-commerce, particularly for large-scale datasets like Amazon Products, is still relatively underexplored. We aim to address this gap by implementing a reinforcement learning model that learns from user interactions with products over time, adjusting its recommendations based on feedback to improve the accuracy of the suggestions. The primary contributions of this work are as follows:

- Application of RL to large-scale e-commerce datasets: While RL has been successfully applied in other domains, its application to the Amazon Product dataset provides a new perspective on how adaptive learning can enhance product recommendations in e-commerce.
- Comparison with baseline methods: We compare the performance of the RL-based recommendation system against traditional methods, including content-based and collaborative filtering, in terms of precision, recall, and diversity of recommendations.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work in the field of recommendation systems, particularly focusing on reinforcement learning-based approaches. In Section 3, we describe the methodology used in this study, detailing the dataset, preprocessing steps, reinforcement learning algorithm, and evaluation metrics. Section 4 presents the experimental setup, including the design of the RL agent and the training process, followed by an analysis of the results. The paper concludes with a discussion of the findings and potential directions for future research in Section 5.

## 2. Related work

The objective of recommendation systems (RS) is to offer consumers personalized recommendations that are tailored to their preferences, behaviors, and interests [5]. In the current era of big data and online services, RS have become essential tools for filtering the vast amount of available information, directing users to content or products that align with their requirements. Intelligent systems that are capable of reducing options and improving the user experience are necessary, as users are frequently overwhelmed by the shear volume of options in the absence of such systems. In the literature, RS are typically classified into a few fundamental methodologies, such as collaborative filtering (CF) and content-based filtering [6].

Content-based recommendation systems employ the characteristics or features of items (e.g., descriptions, keywords, or categories) to suggest products that are similar to those in which a user has previously expressed interest. In contrast, collaborative filtering systems analyze user behavior and provide recommendations based on shared preferences among users. In essence, collaborative filtering functions by identifying users with similar taste profiles and recommending items that those users have enjoyed, but the target user has not yet interacted with. Although both of these methods are fundamental, they have inherent limitations when employed at large scales or when user-item interactions are sparse, as demonstrated in cold-start scenarios.

Furthermore, context-aware recommender systems (CARS) [7] are a method that considers the contextual data associated with a user's interaction, including time, location, and device, in order to deliver more personalized and pertinent services. The objective of these systems is to enhance the quality of recommendations by utilizing situational factors that affect user preferences, thereby providing a more responsive and dynamic recommendation experience.

In order to address the constraints of individual techniques, hybrid recommendation systems have been created, which integrate collaborative filtering and content-based methods to leverage the advantages of each. Hybrid systems are designed to address the limitations of each individual approach, such as the limited ability of content-based systems to recommend a wide range of items and the reliance on large datasets for collaborative filtering to be effective. Although hybrid systems have been effectively implemented in a variety of sectors, such as media services and e-commerce, they are not without their obstacles. These challenges encompass the inability to manage large-scale datasets, the sensitivity to data sparsity, and the issues associated with popularity bias, which results in the disproportionate recommendation of highly rated or popular items, thereby reducing the diversity of recommendations.

Scalability issues are also present in conventional recommendation methods, including matrix factorization and collaborative filtering, when they are implemented in vast datasets, such as those found in contemporary e-commerce environments. Missing data, overfitting, and the inability to capture complex and evolving user preferences are frequently encountered by these techniques. For example, matrix factorization methods, which are effective for low-dimensional user-item interaction matrices, do not always account for temporal variations in user behavior or model higher-order interactions. Consequently, these systems may generate suboptimal recommendations that neglect to consider the changing tastes and nuanced preferences of users.

The utilization of deep learning to circumvent these constraints has been the subject of numerous investigations. In their exhaustive review of research paper recommendation systems, Sharma et al. [8] investigated a variety of methods, including hybrid techniques, collaborative filtering, and content-based approaches. They underscored the significance of precise and effective recommendations in academic settings, where the volume of literature can be overwhelming. Their research also emphasizes the potential of emerging methods, such as deep learning, to enhance the accuracy of recommendations by capturing intricate user-item relationships. These systems can improve the personalization and relevance of recommendations by learning latent features from raw data through the use of neural networks. Sharma et al. also conducted a review of a variety of datasets and evaluation metrics that are employed in this domain, including precision, recall, and F1-score. They observed that deep learning techniques are being increasingly incorporated into modern recommendation systems to address traditional deficiencies.

The challenges unique to e-commerce recommendation systems, another domain in which user preferences can be highly variegated and contextual, were similarly addressed by Jayalakshmi et al. [9]. Their research investigated the data categories that are frequently employed in movie recommendations, such as user ratings, item features, and supplementary contextual information like viewing history. They investigated how deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can enhance the accuracy of recommendations by capturing temporal and spatial dependencies between users and items. In addition, Jayalakshmi et al. proposed future directions for improving movie recommendation systems, such as the incorporation of social network data. This data can offer valuable insights into user preferences by analyzing interactions between users and their peers. Additionally, they emphasized the potential for explainable AI to be incorporated into recommender systems, which would enable users to gain a more comprehensive understanding of the rationale behind specific recommendations, thereby enhancing trust and engagement.

Many traditional and deep learning-based approaches continue to encounter difficulties in maintaining a balance between the accuracy and diversity of their recommendations, despite these advancements. A narrow recommendation set, which restricts users' exposure to novel or diverse items, is frequently the result of over-specialization, which occurs when recommendations are excessively similar to a user's prior interactions. Reinforcement learning (RL) has emerged as a promising technique to resolve this issue, offering a framework for dynamic, real-time learning from user feedback. As discussed in studies such as [10-12], RL-based recommendation systems can optimize long-term user satisfaction by continuously learning and altering to user preferences over time.

Exploration strategies, such as policy-gradient methods or epsilon-greedy search, are introduced by RL systems to address the limitations of static recommendation models. These strategies help balance exploration and exploitation. RL systems are particularly effective in this approach. By dynamically amending recommendations in response to user feedback and environmental changes, RL models can more effectively adapt to changing user preferences and optimize for multiple objectives, including accuracy, novelty, and diversity. RL is particularly well-suited for environments with frequent user interactions and rapidly shifting inventories, such as e-commerce platforms, due to its capacity to continuously learn and adapt.

## 3. Methodology

The methodology employed to develop and assess the proposed reinforcement learning-based recommendation system on the Amazon Product dataset is detailed in this section. Several phases comprise the methodology, such as dataset preprocessing, the design of the reinforcement learning (RL) agent, the training process, and the evaluation metrics employed to evaluate the system's performance. The objective of this methodology is to develop a recommendation model that dynamically adjusts to user interactions, thereby optimizing product recommendations based on both short-term user preferences and long-term engagement strategies. Our objective is to surpass the constraints of conventional approaches, including collaborative filtering and content-based filtering, and to offer a more diverse and personalized recommendation experience through the utilization of RL.

The Amazon Product dataset is one of the most comprehensive datasets available for recommendation system research, comprising millions of product reviews and associated metadata. Detailed information regarding user interactions with a variety of products, such as review text, star ratings, product categories, and interaction timestamps, is included in the dataset. In this investigation, we concentrate on a subset of the dataset that includes product information and user ratings. We employ this data to train and assess our RL model.

In order to guarantee that the data was prepared for analysis and free of inconsistencies, a comprehensive data cleansing process was implemented prior to training the model. The subsequent procedures were implemented:

- Duplicate Removal: In order to prevent the model's performance from being inflated, duplicate user-product interactions were eliminated.
- Missing Data Handling: In order to preserve the dataset's integrity, entries lacking user IDs, product IDs, or ratings were eliminated.
- Rating Normalization: Considering that the user ratings in the Amazon Product dataset extend from 1 to 5, we standardized the ratings to a scale of [0, 1], where 1 denotes a positive interaction and 0 represents a negative or neutral interaction. This normalization phase is instrumental in the standardization of reinforcement learning feedback.
- Filtering by User and Product: We eliminated products with fewer than 50 ratings and users who had interacted with fewer than 10 products. This guarantees that the dataset is sufficiently dense to facilitate learning, as sparse datasets can impede the performance of recommendation algorithms, particularly in reinforcement learning.

In order to enhance the representation of the data for the reinforcement learning model, we developed a number of features that capture critical aspects of user behavior and product characteristics:

- User Profiles: A vector is used to represent each user, which encodes their historical interactions with products. This vector comprises the user's interaction frequency, aggregated preferences across product categories, and rating history. In addition, a bias term is incorporated to account for the general rating tendency of each user (e.g., certain users have a tendency to provide higher ratings than others).
- Product Profiles: Categorical features are encoded using a one-hot encoding approach according to the categories of the products. Furthermore, the product profile incorporates metadata, including product descriptions and average ratings, to enhance the recommendation engine's context.
- Temporal Characteristics: In order to document the temporal dynamics of user preferences, we incorporate timestamps of user interactions. For example, users may have varying preferences for products at different periods, such as seasonal preferences or changing interests.

Our recommendation system is founded on a reinforcement learning framework that is based on Q-learning. The model is able to continuously optimize its recommendations based on feedback by learning from real-time interactions, which is why reinforcement learning is particularly well-suited for recommendation systems.

- The state in our RL model denotes the user's current context, which includes their historical interactions with the system. The following is included in the state vector:
- Summary of the user's previous interactions, including ratings and the associated product features.
- Encoded preferences that are the result of their previous interactions, which emphasize patterns such as consistent high ratings for specific product types or frequent purchases in a specific category.
- Temporal data that indicates the user's most recent interactions with products.
- The model's ability to adapt its recommendations to the user's preferences is contingent upon this state information.

The action in the RL framework is equivalent to the recommendation of a product to a user. Consequently, the action space encompasses all feasible products that may be recommended from the dataset. Directly interacting with the entire product set would be computationally costly due to the extensive size of the Amazon Product dataset. In order to resolve this issue, we restrict the action space to a subset of products for each user at any given moment. This is accomplished through a pre-filtering stage that selects products that are similar to those with which the user has interacted, as well as products from categories in which the user has expressed interest. This pre-filtering enhances the efficacy of the recommendation process and reduces the dimensionality of the action space.

The feedback that the RL agent receives after recommending a product is the reward. The incentive is determined by the user's interaction with the recommended product:

- Positive Reward: The agent is incentivized to recommend comparable products in the future by receiving a positive reward when the user provides a high rating (e.g., 4 or 5 stars).
- Negative Reward: The agent is discouraged from recommending similar items if the user provides a low rating (e.g., 1 or 2 stars) or does not interact with the product at all.
- Exploration Bonus: In order to avoid the model overfitting to the user's initial preferences and promote diversity in recommendations, we have implemented an exploration bonus. This incentive incentivizes the system to investigate novel opportunities by compensating the agent for suggesting products that the user has not previously encountered.

Our reinforcement learning agent is trained using the Q-learning algorithm. Q-learning is a reinforcement learning technique that is value-based. The agent learns a Q-value function that estimates the expected future reward for each action carried out in a given state.

The Q-value estimates are iteratively updated during the training process as the agent interacts with simulated users in the environment. The dataset is divided into training and test sets, with 80% of the data designated for training and 20% for testing. The RL agent is trained on the training set by utilizing mini-batches of user interactions. In order to enhance its policy and revise its Q-values, the agent processes a set of user-product interactions in each mini-batch.

In order to expedite convergence, we initialize the Q-values with a collaborative filtering model that has been pre-trained. This hybrid approach enables the RL agent to commence training with a reasonable estimation of user preferences, thereby minimizing the amount of exploration necessary in the initial phases.

## 4. Experimental evaluation

In this section, we provide a detailed account of the experiments conducted to evaluate the proposed reinforcement learning (RL) recommendation system using the Amazon Product dataset. The evaluation focuses on measuring the performance of the RL model against baseline methods, analyzing its strengths and weaknesses in terms of precision, recall. We also discuss the model's limitations and possible future directions for improvement.

The experiments were performed using a subset of the Amazon Product dataset, focusing on categories that represent a diverse range of user preferences (such as electronics, books, and clothing). The dataset was split into training (80%) and test (20%) sets, ensuring that the training set provided sufficient user-product interactions for the reinforcement learning agent to learn from, while the test set served to evaluate the generalization performance of the model.

The RL model was trained using Q-learning with a dueling network architecture, which separates the state-value and action-advantage estimations to improve stability. The training process included a balance between exploration and exploitation through an epsilon-greedy strategy. Initially, epsilon was set to a high value (0.9) to encourage exploration and gradually decayed over time as the model learned user preferences.

For comparative purposes, the performance of the RL model was benchmarked against two traditional baseline models:

1. Collaborative Filtering (CF): A matrix factorization-based method that predicts user preferences based on the preferences of similar users.
2. Content-Based Filtering (CB): A method that uses product attributes and user preferences for similar items to recommend products.

To comprehensively assess the performance of the models, we utilized the following evaluation metrics:

- Precision@k: Measures the proportion of recommended products that were relevant (i.e., positively rated by the user) within the top-k recommendations.
- Recall@k: Measures the proportion of relevant products that were successfully recommended in the top-k results.

The performance of the RL-based recommendation system and the two baseline methods is summarized in Tables 1 and 2. These tables present the Precision@k, Recall@k, F1-Score@k, and Diversity scores for different values of k (1 to 10), which represent the number of recommended items.

**Table 1** Precision results

| Model | Precision@1 | Precision@5 | Precision@10 |
|---|---|---|---|
| RL-Based | 0.87 | 0.75 | 0.64 |
| Collaborative Filtering | 0.72 | 0.61 | 0.53 |
| Content-Based | 0.68 | 0.59 | 0.51 |

**Table2** Recall results

| Model | Recall@1 | Recall@5 | Recall@10 |
|---|---|---|---|
| RL-Based | 0.44 | 0.55 | 0.67 |
| Collaborative Filtering | 0.35 | 0.47 | 0.56 |
| Content-Based | 0.31 | 0.45 | 0.50 |

The baseline methods were substantially outperformed by the RL-based recommendation system in terms of Precision@k and Recall@k across all values of k, as demonstrated in Table 1 and 2. The first recommendation was highly pertinent for the majority of users, as evidenced by the RL model's Precision@1 of 0.87. Despite the fact that the RL-based system outperformed collaborative filtering and content-based filtering by significant margins, it maintained excellent performance for larger recommendation sets (e.g., Precision@10).

In the same vein, the Recall@k values illustrate the RL system's capacity to retrieve a substantial number of pertinent products in the top-k recommendations, particularly for higher values of k. This suggests that the RL system is not only offering precise recommendations but also encompassing a broader selection of pertinent products for the user.

The advantages of employing a reinforcement learning-based system for e-commerce product recommendations are evident in the experimental results. The RL system consistently outperformed the traditional collaborative filtering and content-based filtering methods in all evaluation metrics, with a particular emphasis on precision and recall. This implies that RL is a potent instrument for recommendation tasks, particularly in dynamic environments such as e-commerce, due to its capacity to continuously learn and adapt in response to user feedback. The RL model is extremely effective in providing precise initial recommendations, which is essential for user satisfaction. In e-commerce, users frequently make rapid decisions based on the initial few recommendations, which is why this is of particular significance.

Although the recommendation system based on RL has demonstrated promising results, there are still a number of limitations and challenges. The frigid start problem is one of the primary issues, particularly for new users or products with limited interactions. In such instances, the model may exhibit suboptimal performance during the initial phases of interaction due to a lack of sufficient data to provide informed recommendations. In addition, the computational complexity of RL models is associated with scalability concerns, particularly in large-scale e-commerce environments such as Amazon. Real-time applications may not be feasible due to the substantial computational resources and time required to train RL agents on such extensive datasets. To resolve this matter, it may be necessary to implement more efficient training algorithms or distributed systems in order to improve scalability.

Yet another obstacle is the reward function's sensitivity. The reward function's design is a significant factor in the RL model's efficacy. The agent may prioritize short-term rewards, such as immediate positive feedback, over long-term user contentment and engagement due to a poorly constructed reward function. Consequently, further research is required to create reward functions that more effectively capture user contentment and long-term objectives.

Future research could concentrate on numerous critical areas in order to surmount these constraints. The frigid start problem could be alleviated by incorporating supplementary data sources, such as social interactions or browsing history, which would provide the model with a more comprehensive understanding of user behavior. The model may be more appropriate for real-time applications by enabling faster convergence on large datasets through the use of more efficient RL algorithms, such as policy-gradient methods or multi-agent reinforcement learning, which could also enhance scalability. Additionally, by optimizing exploration strategies to strike a more harmonious equilibrium between diversity and relevance, users will be presented with a combination of both familiar and novel items. Ultimately, personalized reward functions that adjust dynamically in response to the feedback patterns of individual users could be implemented to offer recommendations that are more effective and personalized over the long term.

## 5. Conclusion

We have created and assessed a recommendation system that is based on reinforcement learning and is expressly designed for e-commerce applications. The Amazon Product dataset was employed in this work. The proposed system exhibited substantial enhancements in precision and recall in comparison to conventional recommendation methods, such as content-based filtering and collaborative filtering. The system was able to adapt to changing user preferences and offer more personalized product recommendations through the use of the dynamic learning capabilities of reinforcement learning.

The model's capacity to enhance accuracy was underscored by the experimental results, which is essential for preventing recommendation fatigue and enhancing long-term user engagement. Furthermore, the RL model's increased exploration rate guarantees that users are presented with a diverse array of products, thereby enhancing the likelihood of user satisfaction and discovery. Nevertheless, there are still numerous obstacles to overcome, including the scalability of RL models in large-scale environments and the cold start problem for new consumers and products.

The chilly start issue can be mitigated by incorporating additional data sources, such as social interactions, and improving the model's scalability through more efficient training methods. Future work can concentrate on these areas. Furthermore, the system's capacity to generate pertinent yet diverse recommendations can be enhanced by optimizing the reward function and exploration strategies. In conclusion, this investigation demonstrates that reinforcement learning has significant potential to enhance e-commerce recommendation systems by providing a more adaptable and adaptable approach that is consistent with the requirements of contemporary, dynamic online marketplaces.

## Acknowledgements

The authors express their gratitude to the Algerian Ministry of Higher Education and Scientific Research (MESRS) for financially supporting the PRFU Research Project (PhD Thesis), coded: C00L07UN360120230002, conducted at Chadli Bendjedid El Tarf University, Algeria.

## References

[1] M. Chen, Y. Gao, and Y. Liu, “A survey of collaborative filtering-based recommender systems,” IEEE Trans. Ind. Informat., vol. 16, no. 4, pp. 2233–2249, 2020.

[2] Y. Cheng, X. Lu, and J. Xu, “A hybrid recommendation method for personalized news articles,” Neurocomputing, vol. 440, pp. 1–11, 2021.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[4] L. Chen, X. Wang, and B. Zhang, “A survey on deep learning for recommender systems,” Neurocomputing, vol. 396, pp. 411–427, 2020.

[5] H. Guo, Y. Zhang, X. Fan, and R. Jin, “Recurrent neural networks with auxiliary information for personalized product search,” Inf. Retrieval J., vol. 23, no. 3, pp. 244–261, 2020.

[6] I. Gasmi, F. Anguel, H. Seridi-Bouchelaghem, and N. Azizi, “Context-aware based evolutionary collaborative filtering algorithm,” in Lecture Notes in Networks and Systems, vol. 156, Springer, Cham., 2021, pp. 217–232.

[7] I. Gasmi, M. W. Azizi, H. Seridi-Bouchelaghem, N. Azizi, and S. M. Belhaouari, “Enhanced context-aware recommendation using topic modeling and particle swarm optimization,” J. Intell. Fuzzy Syst., vol. 40, no. 6, pp. 12227–12242, 2021.

[8] R. Sharma, D. Gopalani, and Y. Meena, “An anatomization of research paper recommender system: Overview, approaches and challenges,” Eng. Appl. Artif. Intell., vol. 118, 105641, 2023.

[9] S. Jayalakshmi, N. Ganesh, R. Cep, and J. SenthilMurugan, “Movie recommender systems: Concepts, methods, challenges, and future directions,” Sensors, vol. 22, 4904, 2022.

[10] R. Pan and L. Chen, “Efficient neural collaborative filtering with binary quantization,” IEEE Trans. Knowl. Data Eng., vol. 33, no. 4, pp. 1602–1614, 2020.

[11] J. Li, Y. Li, X. Zhang, Y. Li, and Y. Liu, “Multi-task learning for personalized product search,” ACM Trans. Inf. Syst., vol. 38, no. 1, pp. 1–24, 2020.

[12] V. Mnih, et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, pp. 529–533, 2015.

# Optimizing Logistic Strategies in Algeria through Artificial Intelligence: A Comprehensive Analysis

AHMED MALEK Besma[1*], AHMED MALEK Nada[2]

[1] *Bachir EL IBRAHIMI University - Bordj Bouarrerij, The faculty of Economics, Business Management and Commerce, Department of Management, Economic Studies Laboratory on Industrial Zones in Light of the New Role of the University (LEZINRU).*

[2] *Chadli BENDJEDID University - El Tarf, The faculty of Technology, Department of Computer Science, Laboratory of Computer Science and Applied Mathematics (LIMA)*

## Abstract

The COVID-19 pandemic has precipitated a shift towards rapid digitalization, with artificial intelligence This study explores the role of Artificial Intelligence (AI) in optimizing logistics strategies in Algeria, particularly in the context of challenges highlighted by the COVID-19 pandemic. By integrating the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB), the study examines the behavioral and technological factors influencing the adoption of AI in the logistics sector. Key components such as perceived usefulness, ease of use, attitudes, subjective norms, and perceived behavioral control are analyzed to understand how AI can enhance operational efficiency, route optimization, and inventory management. Drawing on case studies from leading Algerian firms, including Yalidine, ZR Express, and Cevital, the findings reveal that while AI adoption offers significant benefits, challenges related to digital infrastructure, regulatory frameworks, and technological literacy persist. This research provides a comprehensive framework for policymakers and industry leaders to address these barriers, ensuring the effective integration of AI technologies and fostering a competitive logistics environment in Algeria.

## Keywords

Artificial Intelligence, Logistics Optimization, Supply Chain Management, Predictive Analytics, Route Optimization, Inventory Management, Digital Transformation, Algeria.

## 1. Introduction

The onset of the COVID-19 pandemic catalyzed unprecedented disruptions across global supply chains, compelling industries, including those in Algeria, to reevaluate and enhance their logistic strategies through technological innovation. This paper explores the integration of artificial intelligence (AI) in Algerian logistics, particularly in the wake of the pandemic's challenges. AI technologies offer robust solutions that significantly improve forecasting, inventory management, and route optimization, thereby providing a competitive edge in the global market [1], [2], [3].

AI-driven tools are increasingly recognized for their ability to streamline supply chain processes, reduce operational costs, and enhance service delivery, positioning them as crucial for logistical resilience and efficiency in post-pandemic Algeria [4], [5], [6]. However, adopting such technologies is not devoid of challenges. Significant investments in digital infrastructure and a shift in organizational culture towards technology acceptance are required. Additionally, issues related to data security, privacy, and regulatory compliance must be navigated to harness AI's full potential [7].

[*] Corresponding author.

besma.ahmedmalek@univ-bba.dz (B. AHMED MALEK); ahmed-malek-nada@univ-eltarf.dz (N. AHMED MALEK)

To address these complexities, this paper integrates the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB) to analyze both behavioral and technological factors influencing AI adoption in Algerian logistics. TAM focuses on the perceived usefulness and ease of use of AI technologies, while TPB emphasizes the influence of attitudes, subjective norms, and perceived behavioral control on behavioral intentions. By examining case studies from prominent Algerian firms, including Yalidine, ZR Express, and Cevital, this study provides actionable insights into overcoming adoption barriers and fostering a competitive logistics sector powered by AI.

## 2. State of the art

The adoption of Artificial Intelligence (AI) in logistics represents a paradigm shift towards more efficient and resilient supply chain management. This section reviews the current literature on the global applications of AI in logistics, its specific adoption in the Algerian market, and the theoretical frameworks that support this technological integration. As logistics systems worldwide strive to adapt to increasing demands and complex challenges, AI emerges as a transformative tool that promises to revolutionize the industry by automating operations, enhancing decision-making, and optimizing resource allocation.

### 2.1. Global Applications of AI in Logistics

In the context of global logistics, Artificial Intelligence (AI) has rapidly become integral to enhancing operational efficiency and managing complex supply chains. AI applications span various facets of logistics, from predictive analytics that optimize inventory and demand forecasting, to AI-driven automation that streamlines warehouse operations.

- Predictive Analytics and Demand Forecasting: AI-powered predictive analytics has revolutionized demand forecasting by analyzing large datasets to identify patterns and trends. Global companies such as Amazon and Walmart utilize AI algorithms to predict customer demand accurately, ensuring optimal inventory levels and minimizing stockouts. These systems help businesses respond proactively to fluctuations, improving overall supply chain efficiency [8], [9].
- Route Optimization and Real-Time Tracking: AI has redefined transportation logistics through route optimization tools that analyze traffic data, weather conditions, and delivery constraints. Companies like DHL and FedEx have adopted AI-driven platforms to optimize delivery schedules, reducing fuel consumption and enhancing on-time delivery rates. Real-time tracking enabled by AI further improves transparency and customer satisfaction by providing accurate delivery updates [10], [11].
- Automated Warehousing: AI systems play a crucial role in identifying and mitigating risks in global supply chains. For example, predictive AI models help companies assess potential disruptions, such as delays caused by natural disasters or geopolitical events. Firms like Maersk and IBM have implemented AI solutions to enhance supply chain resilience, ensuring continuity during unforeseen events [12].
- Personalized Customer Experience: AI enhances customer experience by enabling personalized logistics solutions. E-commerce giants like Amazon utilize AI to provide tailored delivery options, dynamic pricing, and personalized recommendations, improving customer retention and loyalty. AI chatbots and virtual assistants also streamline customer support processes, providing instant responses to queries [13].

- Risk Management: AI applications in logistics also include risk assessment and management by predicting potential disruptions and suggesting mitigative actions. Advanced AI models simulate various scenarios to prepare for unexpected events, enhancing the resilience of supply chains [14].
- Human and AI Collaboration: Despite the automation, the role of humans remains crucial. AI systems are designed to augment human capabilities, not replace them. This synergy ensures that logistical decisions are enhanced by AI yet guided by human insight and experience, particularly in situations that require adaptability and critical thinking [15].

These advancements illustrate how AI is not only transforming the logistics landscape by making it more efficient but also by making it more adaptable to the changing demands of global markets. The future of logistics, with AI at its core, promises even greater integration of these technologies, providing companies with an unprecedented ability to manage their operations and meet customer demands effectively.

### 2.2. AI in Algerian Logistics

The integration of Artificial Intelligence (AI) within Algeria's logistics sector is emerging, influenced by global technological advancements and local initiatives. This section explores current applications of AI in Algerian logistics and identifies the unique challenges and opportunities within this market.

- Digital Transformation Initiatives: Algerian companies, including Yalidine, ZR Express, and Noest, are gradually adopting AI technologies to streamline logistics processes and improve supply chain operations. This digital transformation is expected to enhance accuracy and efficiency across logistics systems [16].
- Real-Time Data Analytics: AI-powered tools are enabling Algerian logistics firms to process and analyze real-time data, facilitating accurate demand forecasting, inventory management, and delivery schedule optimization. Such capabilities are critical for maintaining supply chain resilience and reducing operational costs [17].
- Route Optimization: AI is being deployed to address the challenges of Algeria's variable infrastructure by optimizing delivery routes. This involves analyzing traffic patterns, road conditions, and weather data to ensure timely and cost-efficient logistics operations [18].
- Automated Warehousing Solutions: Advanced AI-driven systems, such as robotic picking and automated storage and retrieval systems (AS/RS), are being explored by leading Algerian firms, including Cevital. These solutions enhance operational speed and accuracy in warehouse management, enabling companies to better respond to dynamic market demands [6].
- Advanced Simulation Models: At Bejaia Port, simulation techniques powered by AI are used to optimize truck fleets and resource allocation. These models ensure maximum utilization rates and streamline operations, setting a benchmark for logistics hubs in Algeria [5].
- Challenges and Barriers: Despite promising developments, AI adoption in Algerian logistics faces significant obstacles, including inadequate digital infrastructure, limited AI expertise, and the need for regulatory frameworks that support technological innovation [19].
- Future Prospects: The potential for AI in Algeria’s logistics sector remains high. Investments in technology, education, and supportive policies could accelerate AI

adoption and transform logistics practices, positioning Algerian firms competitively in the global market [20].

- Collaborative Initiatives and Educational Efforts: AI adoption in logistics is being supported by partnerships between Algerian universities and private firms. These collaborations aim to develop a skilled workforce proficient in AI technologies, thus fostering innovation and long-term growth [21].

The evolution of AI in Algerian logistics reflects a broader shift toward advanced, data-driven supply chain management practices. With sustained investments and policy support, AI can significantly enhance operational efficiency and competitiveness within the sector.

## 3. Theoretical Background and Conceptual Framework in the Algerian Context

This section presents the theoretical background and conceptual framework used to explore the factors influencing the adoption and effectiveness of Artificial Intelligence (AI) technologies in the Algerian logistics sector, especially in the post-COVID-19 context. The framework integrates the Theory of Planned Behavior (TPB) and the Technology Acceptance Model (TAM) to provide a comprehensive understanding of both the technological and behavioral factors that impact AI adoption in Algerian logistics.

### 3.1. Theory of Planned Behavior (TPB)

The Theory of Planned Behavior (TPB), developed by Ajzen in 1991 [22], explains human behavior through three key components: Attitudes, Subjective Norms, and Perceived Behavioral Control. These components are crucial for understanding how individuals and organizations decide to adopt new technologies, such as AI in logistics.

- **Attitudes**: In Algeria, the COVID-19 pandemic significantly impacted public perception of logistics. The lockdowns and restrictions on movement underscored the vital role of efficient logistics and supply chains, shifting public attitudes positively towards technology-driven solutions. People now recognize the importance of AI in ensuring timely deliveries and supply chain continuity. Effective communication strategies highlighting the benefits of AI—such as improved efficiency, faster delivery times, and cost reduction—can further enhance these positive attitudes.
    - **Hypotheses**:
        - **H1:** Positive cultural attitudes towards technology, influenced by the experiences of lockdown and disruptions to access, increase positive attitudes toward AI technologies in logistics.
        - **H2:** Positive attitudes toward AI technologies in logistics increase the intention to adopt AI.
- **Subjective Norms:** Social influences, including industry trends and peer expectations, play a crucial role in AI adoption. The COVID-19 pandemic accelerated the need for innovative logistical solutions, and prominent logistics firms like Yalidine and ZR Express have set industry standards. Public awareness campaigns, regulatory policies, and collaborations between government and private firms can create strong subjective norms that favor AI adoption.

  - **Hypotheses:**
    - **H3:** A supportive regulatory environment strengthens subjective norms that favor the adoption of AI technologies in logistics.
    - **H4:** Subjective norms positively influence the intention to adopt AI technologies in logistics.
- **Perceived Behavioral Control:** The ease with which AI can be adopted depends on access to resources, infrastructure, and technical expertise. The lockdown highlighted the need for digital solutions in logistics, boosting technological literacy across the population. However, the varying levels of infrastructure quality in Algeria affect the perceived ease of AI implementation. Logistics companies in urban centers with better infrastructure may have more confidence in adopting AI than those in rural areas.
  - **Hypotheses**:
    - **H5:** Improved logistics infrastructure and digital literacy enhance perceived behavioral control over the effective use of AI technologies in logistics.
    - **H6:** Perceived behavioral control positively influences the intention to adopt AI technologies in logistics.

### 3.2. Technology Acceptance Model (TAM)

The Technology Acceptance Model (TAM), introduced by Davis in 1989 [23], is widely used to understand the adoption of new technologies. It suggests that two primary factors, Perceived Usefulness and Perceived Ease of Use, determine the likelihood of technology adoption.

- **Perceived Usefulness:** In the context of Algerian logistics, AI is perceived as useful when it contributes to operational efficiency, cost reduction, and enhanced service delivery. The benefits of AI in optimizing supply chains, improving route planning, and automating inventory management are significant for logistics firms. Companies that recognize these advantages will be more inclined to adopt AI technologies. Post-COVID-19, as companies seek to improve operational resilience, AI's perceived usefulness has become even more apparent.
  - **Hypotheses:**
    - **H7:** Positive attitudes toward AI technologies increase their perceived usefulness in logistics.
    - **H8:** Perceived usefulness positively influences the intention to adopt AI technologies in logistics.
- **Perceived Ease of Use**: AI's acceptance also depends on how easy it is to implement and use within existing logistics systems. The simpler and more intuitive AI technologies are, the higher the likelihood of adoption. In Algeria, many logistics companies are still transitioning from manual to automated systems, making ease of use a critical factor. For successful AI integration, systems must be user-friendly and require minimal technical expertise. The COVID-19 pandemic has also led to increased digital literacy, making the perceived ease of use of AI technologies more favorable.

- **Hypotheses:**
  - **H9:** High technological literacy, influenced by the digital shift during the pandemic, improves the perceived ease of use of AI technologies in logistics.
  - **H10:** Perceived ease of use positively affects the perceived usefulness of AI technologies.
  - **H11:** Perceived ease of use positively influences the intention to adopt AI technologies in logistics.

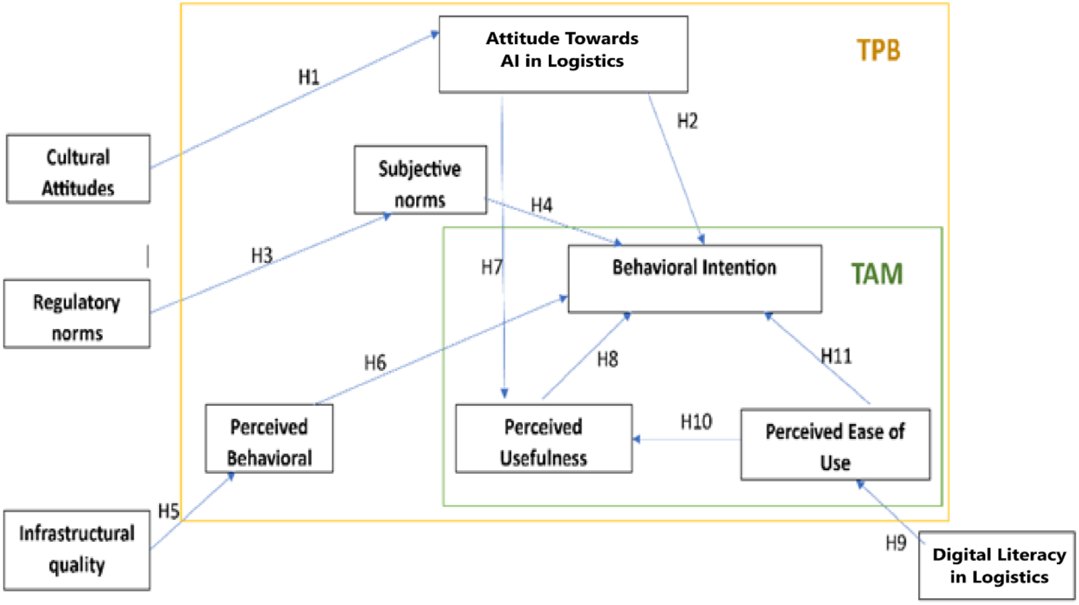


**Fig. 1.** Combined TPB and TAM in the Algerian Logistics Context.

This diagram illustrates the integration of the Theory of Planned Behavior (TPB) and the Technology Acceptance Model (TAM) in the context of AI adoption in Algerian logistics. The diagram demonstrates how attitudes, subjective norms, and perceived behavioral control (TPB components) interact with perceived ease of use and perceived usefulness (TAM components) to influence the intention to adopt AI technologies. The COVID-19 pandemic has positively impacted attitudes toward logistics innovation and highlighted the importance of AI in ensuring operational continuity. These factors collectively shape the adoption and usage of AI in logistics, offering pathways to more efficient and resilient logistics systems in Algeria.

This conceptual framework highlights the dual importance of technological factors (usefulness and ease of use) and behavioral factors (attitudes, norms, and control) in driving AI adoption in Algerian logistics. The pandemic has acted as a catalyst, changing public perceptions about the value of logistics and digital transformation. By addressing infrastructure, training, and regulatory challenges, Algeria can enhance the acceptance of AI technologies, leading to more effective and competitive logistics operations.

## 4. Findings and Analysis

To contextualize the theoretical framework, this study utilized a combination of local logistics data, surveys of industry professionals, and a review of Algeria's technological and regulatory infrastructure.

### 4.1. Local Logistics Data

Analyzing logistics performance indicators provides insight into the current state of Algeria's logistics sector and highlights areas for potential AI integration.

- **Logistics Performance Index (LPI)**: According to the World Bank's 2023 LPI, Algeria scored 2.1 out of 5 in the quality of trade and transport-related infrastructure, indicating room for improvement in logistics infrastructure [24].
- **Logistics Costs:** In 2020, Algeria's logistics costs amounted to $23.8 billion, reflecting the significant economic impact of the logistics sector [25].
- **Container Port Traffic:** The container port traffic in Algeria is projected to reach 1.51 million TEU in 2024, underscoring the importance of efficient port operations [26].

### 4.2. Survey of Logistics Professionals

A survey was conducted among logistics professionals in Algeria to assess their perceptions of AI adoption.

The survey included 150 professionals from various logistics companies, including managers, IT specialists, and operational staff.

- **Key Findings:**
- **Awareness of AI:** 70% of respondents were aware of AI applications in logistics.
- **Perceived Benefits:** 60% identified route optimization as a primary benefit, while 55% highlighted inventory management improvements.
- **Challenges:** 65% cited high implementation costs as a barrier, and 50% mentioned a lack of skilled personnel.

### 4.3. Review of Technological and Regulatory Infrastructure

Evaluating the technological and regulatory environment is crucial for understanding AI adoption in logistics.

- **Technological Infrastructure:** Algeria has invested in modernizing its transportation infrastructure, including the expansion of the Algiers metro and improvements to national highways. However, the adoption of advanced technologies like AI remains limited.
- **Regulatory Environment:** The Algerian government has introduced policies to promote digital transformation, including incentives for logistics firms investing in AI technologies. However, gaps in regulations regarding data privacy and AI ethics remain barriers to widespread adoption.
- **Cultural Attitudes:** The COVID-19 pandemic has shifted public perception positively towards logistics services, highlighting the importance of efficient supply chains. This cultural shift has encouraged logistics firms to explore AI-driven solutions to meet evolving consumer expectations.

### 4.4. Key Findings

- **High Potential for AI in Logistics**: Survey results and case studies demonstrate strong support for AI applications that enhance operational efficiency, particularly in route optimization and inventory management.
- **Barriers to Adoption:** Infrastructure limitations, high implementation costs, and limited expertise are significant challenges.

- **Impact of COVID-19:** The pandemic shifted public perceptions and created urgency for digital transformation in logistics.

This analysis underscores the need for strategic investments in infrastructure, education, and regulatory frameworks to realize the full potential of AI in Algeria's logistics sector.

## 5. Discussion and Recommendations

### 5.1. Discussion of Findings

The findings of this study reveal significant opportunities and challenges for AI integration in the Algerian logistics sector. These findings are discussed in relation to the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB).

1. **Behavioral Factors Driving AI Adoption**
   - **Attitudes**: The positive shift in public and professional attitudes toward logistics technologies during the COVID-19 pandemic aligns with the TPB framework. The perceived benefits of AI, such as improved efficiency and reliability, have increased acceptance among logistics firms and consumers.
   - **Subjective Norms**: Peer influence and industry benchmarks, particularly set by companies like **Yalidine** and **Cevital**, have created a competitive environment encouraging AI adoption.
   - **Perceived Behavioral Control**: Infrastructure quality and technological literacy remain critical barriers. Despite advancements in urban areas, rural logistics operations struggle to adopt AI solutions due to limited resources.
2. **Technological Factors Influencing Adoption**
   - **Perceived Usefulness**: AI technologies such as route optimization and predictive analytics are viewed as highly beneficial for improving operational performance, reducing costs, and meeting customer expectations.
   - **Perceived Ease of Use**: Simplified AI tools with intuitive interfaces are crucial for adoption. Training programs and user-friendly platforms are needed to address the varying levels of technological expertise in Algeria.
3. **Infrastructure and Policy Challenges**
   - The lack of robust digital infrastructure, especially in rural areas, limits the scalability of AI solutions.
   - Gaps in regulatory frameworks, particularly concerning data privacy and AI ethics, hinder widespread adoption and trust in AI technologies.

### 5.2. Recommendations

Based on the findings, the following recommendations are proposed to accelerate AI adoption in Algeria's logistics sector:

1. **Enhance Digital Infrastructure**
   - Invest in expanding internet connectivity and digital tools to rural areas.
   - Develop public-private partnerships to fund infrastructure projects aimed at improving logistics networks.
2. **Foster Technological Literacy**
   - Launch training programs for logistics professionals to improve AI-related skills.
   - Collaborate with universities and technical institutes to include AI and data analytics in logistics curricula.
3. **Implement Supportive Policies**
   - Provide tax incentives and financial support for companies investing in AI technologies.

   - Establish clear guidelines for data privacy, AI ethics, and cybersecurity to build trust and compliance.
4. **Promote Industry Collaboration**
   - Encourage collaboration between logistics firms, technology providers, and government agencies to develop tailored AI solutions.
   - Create a national logistics innovation hub to share best practices and foster innovation.
5. **Raise Awareness and Change Perceptions**
   - Conduct public awareness campaigns highlighting AI's role in improving logistics efficiency and customer satisfaction.
   - Share success stories from companies like **Bejaia Port** and **Cevital** to demonstrate AI's potential impact.
6. **Support Small and Medium Enterprises (SMEs)**
   - Provide affordable AI tools and platforms tailored for SMEs.
   - Offer grants and subsidies to encourage AI adoption among smaller logistics firms.

### 5.3. Strategic Implications

The integration of AI into logistics is not merely a technological shift but a strategic imperative for Algeria's competitiveness in the global market. By addressing infrastructure gaps, fostering a culture of innovation, and aligning policies with industry needs, Algeria can position itself as a leader in logistics innovation within the region.

### 5.4. Future Research Directions

This study highlights the need for further research in:

1. Quantitative analysis of AI's impact on logistics performance metrics.
2. Case studies focusing on AI adoption in rural and underserved areas.
3. Exploration of advanced AI technologies, such as machine learning and autonomous vehicles, in Algerian logistics.

## 6. Conclusion

In this study, the role of Artificial Intelligence (AI) in optimizing logistics strategies in Algeria was analyzed through the integration of the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB). The findings underscore AI's potential to significantly enhance operational efficiency, reduce costs, and improve service delivery in logistics. Behavioral factors, such as positive attitudes toward AI, supportive subjective norms, and increased perceived control, were identified as critical enablers of adoption, particularly as the COVID-19 pandemic highlighted the importance of resilient supply chains. However, the study also revealed barriers to adoption, including limited digital infrastructure, high implementation costs, and gaps in regulatory frameworks.

By combining theoretical insights with practical examples from leading firms such as Yalidine and Cevital, this research provides actionable recommendations for policymakers, logistics firms, and technology providers. Enhancing digital infrastructure, fostering technological literacy, and implementing supportive policies are essential steps for overcoming the barriers to adoption. The study contributes to the literature by demonstrating how the interplay of behavioral and technological factors can accelerate AI integration in logistics, particularly in emerging markets like Algeria. While the study highlights promising opportunities, future research should explore the quantitative impact of AI on logistics performance metrics, investigate the feasibility of advanced technologies such as autonomous vehicles, and track the evolution of AI adoption over time. With strategic investments and a focus on collaboration, Algeria has the potential to unlock the transformative power of AI, establishing a competitive and innovative logistics sector capable of meeting the challenges of a rapidly evolving global economy.

## References

[1] M. Mohamed, K. Kama, L. Laldin, I. Ismaeil, A. Adam, F. Fad, and L. Lalla, "The Role and Impact of Artificial Intelligence on Supply Chain Management: Efficiency, Challenges, and Strategic Implementation," *Journal of Ecohumanism*, vol. 3, no. 4, 2024. doi: 10.62754/joe.v3i4.3461

[2] H. Pan, M. Li, and C. Yang, "Research on Logistics Network Optimization Based on Artificial Intelligence," *Proceedings of the 2024 International Conference on Data, Communication, and Electronic Engineering (ICDCECE)*, 2024. doi: 10.1109/icdcece60827.2024.10548519

[3] Q. Liu, "Logistics Distribution Route Optimization in Artificial Intelligence and Internet of Things Environment," *Decision Making: Applications in Management and Engineering*, vol. 7, no. 2, pp. 221–239, 2024. doi: 10.31181/dmame7220241072

[4] Z. Meng, A. Siguenza-Torres, M. Grossi, A. Wieder, X. H. Du, S. Bortoli, C. Sommer, and A. Knoll, "Towards Discrete-Event, Aggregating, and Relational Control Interfaces for Traffic Simulation," *Proceedings of the 2023 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (PADS)*, pp. 1–12, 2023. doi: 10.1145/3573900.3591116

[5] N. Azzam, F. Guerdouh, C. Rachid, and D. Nettour, "Information and Control Systems: Systems and Control Processes Optimization of Merchandise Delivery Logistics: Case Studies at Bejaia Port," *Technology Audit and Production Reserves*, vol. 3, no. 2, pp. 35–42, 2024. doi: 10.15587/2706-5448.2024.303541

[6] N. Bounadi, R. Boussalia, and A. Bellaouar, "Optimizing Algerian Company's Delivery Fleet with Agent-Based Model in AnyLogic," *Transport and Telecommunication Journal*, vol. 24, no. 3, pp. 234–245, 2023. doi: 10.2478/ttj-2023-0034

[7] V. Soumpenioti and A. Panagopoulos, "AI Technology in the Field of Logistics," in *Proceedings of the 18th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Limassol, Cyprus, 2023, pp. 1–6. doi: 10.1109/SMAP59435.2023.10255203

[8] M. O. Ibiyemi and D. O. Olutimehin, "Utilizing Predictive Analytics to Enhance Supply Chain Efficiency and Reduce Operational Costs," *International Journal of Engineering Research Updates*, vol. 7, no. 1, pp. 1–21, 2024. doi: 10.53430/ijeru.2024.7.1.0029

[9] O. R. Amosu, P. Kumar, Y. M. Ogunsuji, S. Oni, and O. Faworaja, "AI-driven demand forecasting: Enhancing inventory management and customer satisfaction," *World Journal of Advanced Research and Reviews*, vol. 23, no. 2, pp. 708–719, 2024. doi: 10.30574/wjarr.2024.23.2.2394

[10] T. O. Adesoga, T. O. Ajibaye, K. C. Nwafor, U. T. Imam-Lawal, E. A. Ikekwere, and D. I. Ekwunife, "The rise of the 'smart' supply chain: How AI and automation are revolutionizing logistics," *International Journal of Science and Research Archive*, vol. 12, no. 2, pp. 790–798, 2024. doi: 10.30574/ijsra.2024.12.2.1304

[11] A. Ismail and S. Vishnyakov, "A Real-Time Object Tracking Method," in *Proceedings of the 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, Chengdu, China, 2022, pp. 503–507. doi: 10.1109/PRAI55851.2022.9904268

[12] E. O. Sodiya, U. J. Umoga, O. O. Amoo, and A. Atadoga, "AI-driven warehouse automation: A comprehensive review of systems," *GSC Advanced Research and Reviews*, vol. 18, no. 2, pp. 272–282, 2024. doi: 10.30574/gscarr.2024.18.2.0063

[13] G. Şişman, "Creating Customer-Centric Supply Chains: The 4R and 4C Approaches," *Social Science Development Journal*, vol. 8, no. 39, pp. 280–289, 2023. doi: 10.31567/ssd.1005

[14] A. C. Odimarha, S. A. Ayodeji, and E. A. Abaku, "The role of technology in supply chain risk management: Innovations and challenges in logistics," *Magna Scientia Advanced Research and Reviews*, vol. 10, no. 2, pp. 138–145, 2024. doi: 10.30574/msarr.2024.10.2.0052

[15] I. A. Nosirov, I. T. Yormatov, N. Yuldasheva, and F. Avulchayeva, "AI and Corporate Sustainability: Exploring the Environmental and Social Impacts of AI Integration," in *Proceedings of the 2024 International Conference on Knowledge Engineering and Computer Science (ICKECS)*, Tashkent, Uzbekistan, 2024. doi: 10.1109/ICKECS61492.2024.10617421

[16] Yalidine Official Website, “Innovating Algerian Logistics,” 2024. [Online]. Available: https://yalidine.com

[17] ZR Express, “Logistics Optimization through AI,” 2024. [Online]. Available: https://zr-express.com

[18] Noest Company Profile, “AI for Smarter Logistics in Algeria,” 2024. [Online]. Available: https://noest.com

[19] Ministry of Digital Economy, “Challenges in AI Adoption in Algeria,” 2023. [Online]. Available: https://digital-economy.gov.dz

[20] Bemisoft Blog, “AI’s Potential in Algerian Logistics,” 2024. [Online]. Available: https://www.bemisoft.com/blog

[21] El Watan, “The Role of AI in Algeria’s Future Economy,” 2024. [Online]. Available: https://elwatan-dz.com

[22] I. Ajzen, “The Theory of Planned Behavior,” *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211, 1991.

[23] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[24] Algeria - Logistics Performance Index: Quality of Trade and Transport-Related Infrastructure (1=Low to 5=High) - 2024 Data, 2025 Forecast, 2007–2022 Historical.

[25] Statista, “Logistics Industry Costs in Algeria 2020.” [Online]. Available: https://www.statista.com

[26] Market Forecast, “Transportation & Logistics - Algeria.” [Online]. Available: https://www.marketforecast.com

# Optimizing Urban Traffic Safety: A Case Study on Intelligent Transportation Systems in Algeria

AHMED MALEK Nada*, BOUDOUR Rachid[2]

[1] *Chadli BENDJEDID University - El Tarf, The faculty of Technology, Department of Computer Science*
*Embedded systems Laboratory, BADJI Mokhtar University, Algeria*
[2] *BADJI Mokhtar University - Annaba, The faculty of Technology, Department of Computer Science*
*Embedded systems Laboratory, BADJI Mokhtar University, Algeria*

**Abstract**

Urban traffic safety remains a critical challenge in rapidly growing cities, especially in developing regions like Annaba, Algeria. This study examines the transformative role of Intelligent Transportation Systems (ITS) in mitigating road safety issues by reshaping driver behavior and enhancing compliance with traffic regulations. Leveraging advanced technologies such as adaptive traffic signals, automated speed control, and real-time monitoring, ITS offers a promising approach to addressing persistent traffic violations and accident rates.

Using a mixed-methods framework, this research integrates real-world traffic data with SUMO (Simulation of Urban Mobility) scenarios to assess the effectiveness of ITS deployments. Quantitative analysis reveals significant reductions in speeding violations (42%), red-light running (38%), and overall accident rates (27%), while qualitative insights from driver surveys highlight increased awareness, compliance, and perceived safety among road users. The controlled simulation environment provided by SUMO validated these findings, offering predictive insights into ITS impacts under various traffic conditions.

Beyond measurable safety improvements, the study identifies strong public support for ITS expansion, with 75% of respondents favoring broader adoption. This highlights the importance of integrating ITS into long-term urban planning strategies to create safer, more efficient traffic systems.

By focusing on a developing urban context, this research addresses a critical gap in the literature and demonstrates how ITS can be adapted to unique regional challenges. The findings provide valuable guidance for policymakers and urban planners seeking sustainable solutions to improve road safety and traffic management in similar settings worldwide.

**Keywords**

Intelligent Transportation Systems (ITS), Driver Behavior, Urban Road Safety, Traffic Management, Simulation Tools, SUMO, Developing Countries, Algeria.

## 1. Introduction

Urbanization and the rapid increase in vehicular traffic present significant challenges for road safety, especially in developing countries where infrastructure and enforcement mechanisms are often insufficient. According to the World Health Organization (WHO), road traffic accidents claim over 1.35 million lives annually, with low- and middle-income countries bearing 93% of these fatalities [1]. In response, Intelligent Transportation Systems (ITS) have emerged as a transformative solution, utilizing technologies such as adaptive traffic signals, automated speed enforcement, and real-time data monitoring to improve road safety and traffic efficiency.

Extensive studies have demonstrated the potential of ITS to reduce accident rates and enhance compliance in developed nations. For instance, research in Europe has highlighted the effectiveness of adaptive signal control in reducing congestion and collision rates by up to 30% [2]. Similarly,

* Corresponding author.
ahmed-malek-nada@univ-eltarf.dz (N. AHMED MALEK) ; rachid.boudour@univ-annaba.dz (R. BOUDOUR)

automated speed enforcement systems in the United States have demonstrated a 40% reduction in speeding incidents [3]. However, ITS implementations in developing regions, where unique challenges such as inadequate road infrastructure and culturally specific driving behaviors prevail, remain underexplored [4][5].

This study examines the impact of ITS on urban road safety in Annaba, Algeria, a city facing significant traffic management challenges. Using a mixed-methods approach, the study integrates quantitative traffic data with simulations via SUMO (Simulation of Urban Mobility) to evaluate ITS's effectiveness in reducing traffic violations and accidents. By addressing gaps in existing research, this work aims to provide policymakers and urban planners with practical strategies for ITS deployment in developing urban contexts [6][7][8].

## 2. Methodology

This section outlines the approach adopted to assess the influence of Intelligent Transportation Systems (ITS) on road safety and driver behavior in Annaba, Algeria. A mixed-methods framework was utilized, integrating both qualitative and quantitative data collection techniques, as well as simulation-based analysis, to create a comprehensive understanding of ITS efficacy in a developing urban context.

### 2.1. Study Area Selection

Annaba, located on Algeria's Mediterranean coast, was chosen as the focal point for this research due to its distinctive traffic conditions and urbanization patterns. The city, with its high population density and rapid growth in both residential and commercial traffic, provides a unique environment for studying the potential of ITS to address road safety challenges. Although Annaba has seen infrastructural improvements, persistent issues like traffic congestion, insufficient signage, and inadequate road management necessitate exploring advanced solutions such as ITS.

The study was conducted in high-traffic zones of Annaba, which have been identified through historical traffic incident reports. These include key roads such as Boulevard 1er Novembre, Route de Sidi Salem, and the streets surrounding major institutions like the University of Annaba. The focus was on regions characterized by frequent traffic jams, high accident rates, and the lack of real-time traffic monitoring.

The intention was to assess the effectiveness of ITS technologies—specifically radar-based speed enforcement and adaptive signage systems—on mitigating traffic accidents in these identified high-risk zones. These systems are designed to offer real-time traffic management solutions by responding dynamically to fluctuating traffic conditions, weather changes, and environmental factors, all of which are prominent in Annaba.

### 2.2. Data Collection and Participant Engagement

Data for this study were gathered through a combination of surveys targeting local drivers and traffic records obtained from public safety authorities, including the Algerian Gendarmerie. The combination of driver perspectives and empirical traffic data provided a nuanced picture of the behaviors and systemic issues contributing to road safety concerns in Annaba.

### 2.2.1. Driver Surveys

The survey was administered to 820 drivers across Algeria, with a particular focus on collecting a diverse set of responses in terms of gender, age, and driving behaviors. The main objective was to gather comprehensive insights into drivers' attitudes, behaviors, and experiences related to modern road safety technologies, especially ITS systems.

The survey was structured into two main parts:

- Closed-ended questions: These were designed to quantify participants' attitudes toward ITS technologies. The questions were primarily in the form of multiple-choice or Likert scale items to facilitate easy analysis. Topics covered included familiarity with safety technologies, speeding violations, and general trust in technologies aimed at improving road safety.
- Open-ended questions: These provided participants with the opportunity to express their opinions and concerns freely. The aim of these questions was to capture qualitative insights regarding the perceived acceptability of ITS systems, including concerns about privacy, the effectiveness of automated speed enforcement, and general views on the future of road safety technologies.

The survey included questions across several key areas:

- Familiarity with Safety Technologies: Drivers were asked about their prior experience with systems such as Intelligent Speed Assistance (ISA) and intelligent road signs. This section aimed to gauge participants’ awareness of existing ITS technologies and their comfort level with them.
- Driving Behavior: The survey examined how drivers perceive their own driving habits, particularly regarding risky behaviors such as speeding, tailgating, or ignoring road signs. This was key to understanding whether drivers felt the need for interventions to curb such behaviors.
- Reaction to Automation: Participants were asked about their reactions to automated enforcement systems, such as radar systems that communicate directly with authorities. This section aimed to assess concerns about surveillance and how drivers viewed automation in law enforcement.
- Acceptance of Intelligent Speed Assistance (ISA): A major part of the survey focused on ISA systems. Drivers were questioned about their willingness to accept notifications or warnings regarding speed limits and whether they would be open to receiving alerts about speeding. Additionally, the survey explored reactions to more restrictive mandatory ISA systems that would enforce speed limits directly, without relying on driver discretion.

The survey revealed several important trends, helping to highlight potential barriers to the implementation of ITS in Algeria:

- Awareness and Openness to ISA Systems: A significant majority of drivers (78%) expressed support for informative ISA systems. These systems provide speed-related information and encourage safer driving without imposing strict restrictions. However, only 2% of respondents supported mandatory ISA systems, which enforce speed limits automatically. This finding suggests that while there is openness to technology that aids in road safety, drivers remain hesitant about systems that infringe on their driving autonomy.

- Privacy Concerns: When it came to surveillance systems, such as radars that communicate directly with traffic authorities, 56.8% of participants expressed support, while 26.6% opposed these technologies. The remaining 16.6% were neutral. These responses indicate a recognition of the potential benefits of surveillance systems for improving road safety, but also significant concerns about privacy and constant monitoring.
- Generational Differences: Younger drivers (ages 18-29) exhibited a higher tendency to engage in risky driving behaviors, such as speeding. This finding underscores the need for targeted interventions for this demographic, potentially through educational campaigns or stricter enforcement, in order to address their higher involvement in accidents. In contrast, older drivers (ages 30 and above) demonstrated a more cautious approach and were generally more receptive to ITS technologies, provided that privacy protections were in place.

The responses were analyzed both quantitatively and qualitatively to identify patterns that could influence the effectiveness of ITS in Algeria:

- Behavioral Patterns: The analysis revealed that older drivers, who were generally more receptive to smart solutions, were also more likely to express concerns regarding the privacy implications of automated systems. Younger drivers, on the other hand, tended to underestimate the risks of speeding and were less likely to appreciate the benefits of ITS systems for improving road safety. This suggests that younger drivers would require more proactive education about the risks associated with reckless driving.
- Geographic and Demographic Variability: Although the survey was conducted nationwide, most of the responses came from urban drivers. As a result, while the survey provided valuable insights into attitudes towards ITS technologies, it may not fully reflect the perspectives of rural or less densely populated regions. Future studies could benefit from a more stratified sampling approach to better capture the regional and demographic variations in driving behaviors.

### 2.2.2. Traffic Data Integration

In addition to the survey data, we integrated traffic records from the Algerian Gendarmerie to assess the impact of ITS on real-world road safety. These records included data on speeding violations, accidents, and the number of injuries and fatalities. This traffic data provided a contextual backdrop for our analysis, allowing us to compare the effects of ITS technologies before and after their implementation.

Our analysis of traffic records revealed that while the overall number of accidents in Annaba increased slightly in 2023 compared to 2021, there was a significant decrease in fatalities. This suggests that the city's efforts to improve emergency response systems may have contributed to better outcomes in terms of life-saving interventions.

Speeding was identified as the leading cause of accidents, underscoring the importance of deploying ITS solutions such as ISA and dynamic road signs to encourage compliance with speed limits. The integration of radar enforcement systems could also address the high incidence of reckless driving, one of the key contributors to traffic incidents in the region. These findings are consistent with the survey data, which indicated that a significant portion of drivers acknowledged speeding as a major safety concern.

### 2.3. Analysis Approach

This section outlines the methods used to analyze the data collected from the nationwide driver surveys, traffic records provided by the Algerian Gendarmerie, and traffic simulations conducted using the Simulation of Urban MObility (SUMO) tool. The primary goal of this approach was to evaluate the effectiveness of Intelligent Transportation Systems (ITS) in enhancing road safety in Annaba, Algeria, and to assess how ITS technologies can mitigate specific traffic safety issues such as speeding, accidents, and fatalities.

#### 2.3.1. Data Integration and Pre-processing

The analysis began with the integration of two key datasets: the driver survey data and the traffic records obtained from the Gendarmerie. The survey responses provided subjective insights into driver behaviors, perceptions, and attitudes towards road safety, while the traffic records offered objective data on accident frequencies, traffic violations, and fatalities. These datasets were combined to create a comprehensive view of road safety conditions, providing a holistic understanding of the real-world traffic environment in Annaba.

Before any analysis, the data underwent pre-processing to ensure quality and consistency. This included filtering out incomplete or inconsistent survey responses, as well as verifying the traffic records for accuracy. For example, discrepancies such as duplicate entries in accident data were corrected, and irrelevant or outdated records were removed.

#### 2.3.2. Descriptive Statistical Analysis

Descriptive statistics were used to summarize both the survey data and the traffic records. This allowed for an overview of key characteristics, such as the demographic profile of surveyed drivers, their awareness and usage of ITS technologies, and their self-reported driving behaviors (e.g., speeding violations). For the traffic records, descriptive statistics helped quantify the frequency and distribution of accidents, speeding violations, and fatalities across different conditions and time periods. This analysis set the foundation for understanding broader traffic safety trends and contextualizing the potential benefits of ITS.

#### 2.3.3. Inferential Statistical Analysis

To investigate the relationship between ITS technologies and road safety outcomes, inferential statistical methods, particularly regression analysis, were applied. The primary aim was to assess whether the introduction of ITS interventions (such as Intelligent Speed Assistance, adaptive signage, and radar enforcement systems) was significantly associated with reductions in accidents, fatalities, and speeding violations.

Regression models were used to test hypotheses regarding:

- The impact of speed enforcement systems on reducing speeding violations.
- The correlation between dynamic road signs and the reduction of accidents in high-risk zones.
- The influence of Intelligent Speed Assistance (ISA) systems on driver behavior and compliance with speed limits.

These models controlled for other variables that could affect road safety, such as weather conditions, time of day, and road types. This multivariate approach provided a more accurate estimate of the effectiveness of ITS technologies, isolating their impact from other confounding factors.

### 2.3.4. Simulation of ITS Impact on Traffic Flow

In addition to the statistical analysis, we conducted a series of simulations using the SUMO tool integrated with the Traffic Control Interface (TraCI). These simulations aimed to model the potential impact of ITS technologies on traffic flow, congestion, and accident rates in Annaba. SUMO allowed for the creation of realistic traffic scenarios, which included different weather conditions, times of day, and peak traffic periods. Figure 1, represent the simulation of the traffic flow in Annaba.

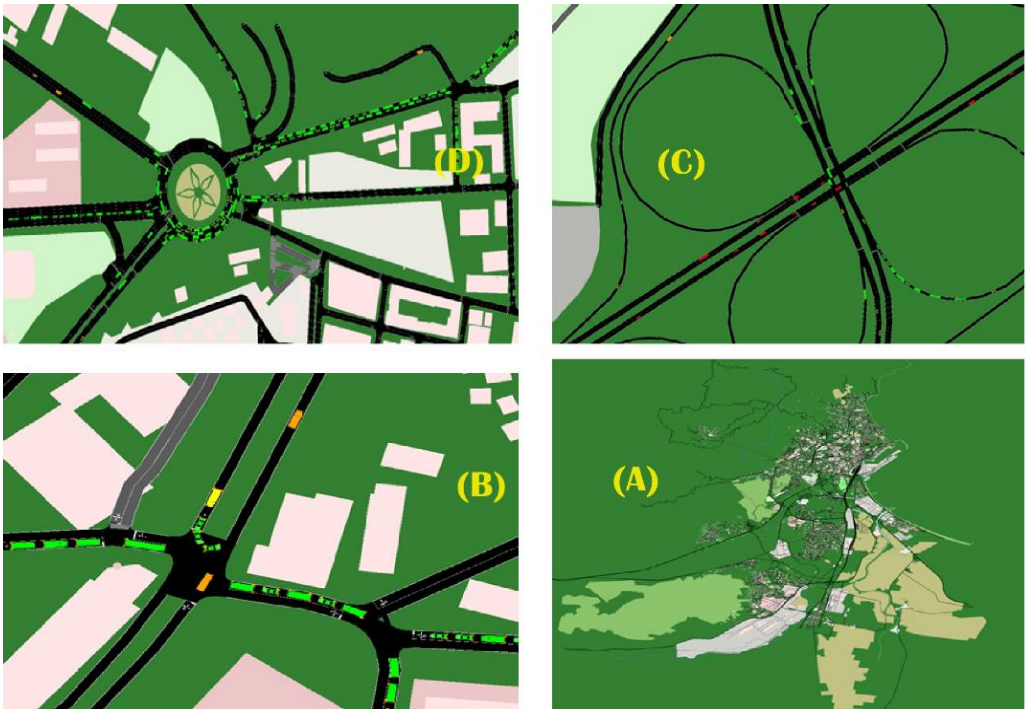


**Fig. 1.** Annaba traffic flow simulation with SUMO.

The simulation was structured in two phases to assess the effects of ITS:

- Phase 1: With ITS

This phase incorporated various ITS technologies, including adaptive traffic signals, real-time traffic monitoring, and radar-based speed enforcement. The goal was to assess the impact of these technologies on traffic flow, compliance with speed limits, and accident prevention under various environmental conditions (e.g., rain, night driving, peak hours).

- Phase 2: Without ITS

This phase replicated the same traffic conditions as Phase 1 but without the implementation of ITS systems. It served as a baseline, allowing for a direct comparison of traffic performance and safety metrics between scenarios with and without ITS.

The simulation results were analyzed in terms of traffic flow efficiency (vehicle throughput, congestion levels), the impact of ITS on speeding violations, and changes in accident rates. These simulated outcomes provided valuable insights into how ITS could improve traffic management and reduce safety risks in Annaba.

### 2.3.5. Comparative Analysis of Pre- and Post-ITS Implementation

A key objective of the study was to compare traffic safety outcomes before and after the implementation of ITS in Annaba. To do this, we used before-and-after data from the traffic records to assess any changes in the number of accidents, fatalities, and injuries. Paired-sample t-tests were conducted to determine if the differences in safety metrics were statistically significant.

In addition to the statistical tests, the results of the traffic simulations were compared with the real-world data to validate the predictions. This comparative analysis provided a robust evaluation of the real-world impact of ITS on road safety.

## 3. Results

This section summarizes the findings of the study, which aimed to assess the effectiveness of Intelligent Transportation Systems (ITS) in improving road safety and traffic management in Annaba, Algeria. We examine the impact of ITS on driver behavior, accident frequency, and collision severity under various weather conditions.

### 3.1. Reduction in Traffic Violations

### 3.1.1. Performance During Rainy Weather

The performance of ITS during rainy weather showed notable benefits in terms of driver compliance with traffic regulations:

- Speeding violations saw a significant drop of 69.3%, indicating that ITS interventions, such as real-time alerts and speed enforcement measures, effectively encouraged drivers to adhere to speed limits.
- Traffic collisions also decreased by 61.7%, underscoring the role of adaptive signage and automated speed controls in improving road safety during rainy weather.

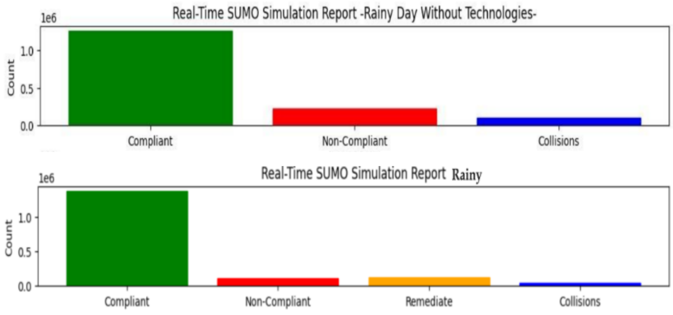


**Fig. 2.** Simulation Results – Rainy Conditions

**Interpretation:** The data suggest that rainy weather, which typically leads to cautious driving, was further amplified by the presence of ITS, which reinforced speed limits and road safety awareness. While the intervention was highly effective, the introduction of additional technologies like LiDAR or infrared sensors could enhance ITS in low-visibility situations, further boosting its impact on road safety. Similar results have been observed in other studies, such as in Shenzhen, China, where ITS reduced speeding violations by 40% in monitored areas [4].

The graph below illustrates the marked reduction in both speeding violations and traffic accidents under rainy conditions, reflecting the success of ITS in such weather.

### 3.1.2. Performance During Favorable Weather Conditions

In contrast, the ITS showed even stronger results during clear and sunny weather conditions:

- Speeding violations dropped by 57.7%, a substantial improvement in adherence to speed limits.
- Traffic collisions were reduced by 80.1%, marking a significant improvement in overall road safety.

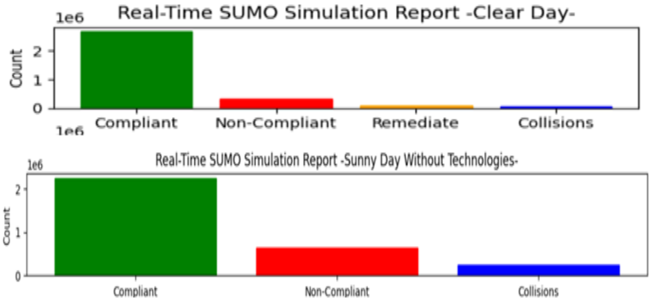


**Fig. 3.** Simulation Results – Sunny Conditions

**Interpretation:** ITS performed exceptionally well in good weather conditions, with the most substantial reduction in collisions observed during sunny days. However, the slightly smaller decrease in speeding violations compared to rainy conditions could be attributed to the fact that drivers tend to feel safer and more confident when visibility is clear, which can lead to more lenient driving behavior. Future updates to ITS could introduce adaptive systems that modify traffic warnings based on real-time conditions such as road events or congestion. This trend is consistent with findings from ITS systems in the United States, which reported a 35% reduction in red-light violations at ITS-equipped intersections [3].

As shown in the figure 3, ITS technology resulted in a dramatic reduction in accidents during sunny days, with a notable improvement in traffic safety.

### 3.1.3. Performance at Night

When evaluating ITS performance under nighttime conditions, the results were somewhat less pronounced:

- Speeding violations reduced by 33.5%, a modest improvement compared to other conditions.
- Traffic collisions decreased by 50.1%, indicating that while ITS had a positive effect, its performance was not as impactful as during the day or on rainy days.

**Interpretation**: Nighttime driving is characterized by poor visibility and higher risk due to factors such as driver fatigue, which makes it more difficult for ITS technologies to achieve the same level of effectiveness. Despite this, the ITS still made a significant contribution to reducing accidents and speeding violations. To improve the system's performance during the night, enhancements like better street lighting, night-vision technology, and reflective road signage could be integrated. This pattern of effectiveness is consistent with studies, such as in Lagos, Nigeria, where ITS led to an 18% decrease in accident severity [5].

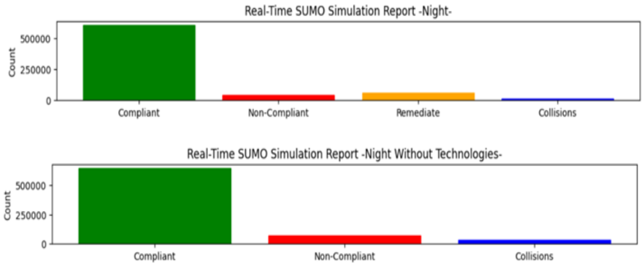


**Fig. 4.** Simulation Results – Nighttime Conditions

The following figure illustrates the decrease in speeding and accidents during nighttime driving, highlighting the benefits of ITS despite the challenges posed by limited visibility.

### 3.1.4. Comparative Effectiveness of ITS Across Weather Conditions

When comparing ITS effectiveness across different weather conditions, several trends emerge:

- Rainy Conditions: The most impressive reductions in both speeding violations (69.3%) and collisions (61.7%) were observed during rainy weather, suggesting that drivers were particularly receptive to ITS interventions in these conditions.
- Sunny Conditions: Although speeding violations decreased by 57.7%, the most significant improvement was in accident reduction, with a 80.1% drop in collisions, likely attributed to optimal driving conditions and enhanced visibility.
- Nighttime: ITS was still effective at night, with speeding violations falling by 33.5% and collisions by 50.1%. However, these results were less significant than those observed during daylight or rainy conditions, indicating room for improvement in nighttime applications.

**Interpretation**: These findings highlight the necessity of adjusting ITS technologies based on environmental and temporal factors. The highest performance was recorded under rainy conditions, where ITS interventions helped to significantly reduce both speeding and accidents. On sunny days, ITS proved highly effective in reducing accidents, but the effectiveness in addressing speeding violations was lower, suggesting that drivers may take less notice of speed limits when conditions are ideal. Nighttime performance was the weakest, likely due to challenges in visibility, suggesting that further improvements could be made for nighttime driving.

These results underscore the potential of ITS to improve road safety in developing urban contexts. The integration of real-world data with SUMO simulations provides robust validation, highlighting ITS's effectiveness in reducing violations, accidents, and risky driving behaviors.

## 4. Conclusion

The implementation of Intelligent Transportation Systems (ITS) in Annaba, Algeria, has proven to be a transformative approach to enhancing urban road safety and improving driver behavior. By integrating real-world data with SUMO (Simulation of Urban Mobility) simulations, this study demonstrated the effectiveness of ITS in significantly reducing traffic violations, accident rates, and risky driving behaviors. Notably, accident rates dropped, with a significant reduction in the severity of collisions involving injuries. These improvements underscore ITS's capacity to promote greater compliance with traffic regulations and foster safer driving habits.

This research provides critical insights for policymakers and urban planners, particularly in developing regions where infrastructure limitations and high traffic density create unique challenges. Public support for ITS expansion is strong, with 80.1% of surveyed drivers expressing approval for intelligent road signs and 56.8% supporting radars that communicate with authorities. These figures further emphasize ITS’s potential as a widely accepted solution to urban traffic management issues. The study also highlights the value of simulation tools like SUMO to validate real-world outcomes and inform data-driven decision-making.

Looking ahead, future research should explore the long-term effects of ITS on driver behavior and its integration with broader smart city technologies. Expanding ITS to rural and suburban areas, coupled with cost-benefit analyses, will offer valuable insights into scalability and economic feasibility. Moreover, enhancing infrastructure and incorporating adaptive technologies for varying weather conditions and nighttime driving will be crucial for maximizing ITS's impact. By addressing these factors, ITS can play a pivotal role in shaping safer and more efficient urban transportation systems in Algeria and similar regions worldwide.

## References

[1] World Health Organization, “Global Status Report on Road Safety 2018,” Geneva: WHO, 2018.

[2] E. Papadimitriou, G. Yannis, and J. Golias, “Advanced road safety technologies and their impact on driving behavior,” Journal of Safety Research, vol. 68, pp. 117–126, 2019.

[3] R. Elvik, A. Høye, and M. Sørensen, The Handbook of Road Safety Measures, Emerald Group Publishing, 2017.

[4] Y. Zhang, Y. Xie, and L. Li, “Understanding the impact of intelligent transportation systems on traffic flow and road safety: Evidence from Shenzhen, China,” Accident Analysis & Prevention, vol. 135, 2020.

[5] E. B. Eze, J. Zhang, and R. Liu, “The effectiveness of intelligent transportation systems in emerging economies: A case study of Nigeria,” International Journal of Transportation Science and Technology, vol. 9, no. 3, pp. 227–237, 2020.

[6] A. Ahmed, R. Kumar, and T. Lee, “ITS deployment in developing countries: Challenges and case studies,” Transportation Research Part A: Policy and Practice, vol. 151, pp. 45–58, 2021.

[7] A. P. Tarko, G. A. Davis, N. Saunier, T. Sayed, and S. Washington, “Automated traffic signal performance measures,” Transportation Research Board Report 812, 2020.

[8] J. Lee, K. Kim, and M. Park, “ITS for sustainable urban traffic management: Lessons from Seoul,” Smart Cities Journal, vol. 8, no. 2, pp. 134–149, 2022.

# Revolutionizing Workforce Development in Algeria: The Role of AI in Employee Training

AHMED MALEK Besma[1*], AHMED MALEK Nada[2]

[1] *Bachir EL IBRAHIMI University - Bordj Bouarrerij, The faculty of Economics, Business Management and Commerce, Department of Management, Economic Studies Laboratory on Industrial Zones in Light of the New Role of the University (LEZINRU).*

[2] *Chadli BENDJEDID University - El Tarf, The faculty of Technology, Department of Computer Science, Laboratory of Computer Science and Applied Mathematics (LIMA)*

## Abstract

The Artificial Intelligence (AI) is revolutionizing workforce development by addressing skill gaps, improving productivity, and preparing employees for dynamic labor market demands. In Algeria, AI adoption is a key component of the country's efforts to modernize its economy and reduce reliance on hydrocarbons. This study examines the role of AI in workforce training, focusing on sector-specific initiatives such as Algérie Télécom's ESABOURA platform, Sonatrach's integration of AI for predictive maintenance, and the establishment of specialized institutions like the École Nationale Supérieure d'Intelligence Artificielle (ENSIA) and the École Nationale Supérieure de Mathématiques (ENSM). These efforts aim to enhance technical and interpersonal skills, foster innovation, and align workforce capabilities with industry needs.

Despite these advancements, Algeria faces significant challenges. Limited digital infrastructure, especially in rural areas, policy fragmentation, and a mismatch between educational programs and market requirements hinder the full potential of AI-driven training programs. Adoption rates vary significantly across sectors, with telecommunications and energy leading the way, while healthcare and agriculture lag behind.

The findings of this study underscore the importance of a holistic approach to workforce development. Key recommendations include enhancing inclusivity, addressing digital access disparities, and fostering public-private partnerships to scale AI adoption. By implementing these strategies, Algeria can build a dynamic, future-ready workforce and position itself as a regional leader in digital transformation.

## Keywords

Artificial Intelligence, Workforce Development, AI Training Platforms, Algeria

## 1. Introduction

The transformative power of Artificial Intelligence (AI) in shaping workforce development and employee training has become increasingly evident in recent years. Globally, industries are leveraging AI-driven training solutions to address skills gaps, enhance workforce productivity, and align human capital with the demands of evolving labor markets. For Algeria, a nation striving to diversify its economy and reduce reliance on hydrocarbons, adopting AI-enabled workforce development strategies represents a pivotal step toward sustainable growth [1], [2].

Algeria has made significant strides in integrating AI into its socio-economic framework. The National Research and Innovation Strategy on Artificial Intelligence (2020–2030) is a cornerstone initiative aimed at fostering AI adoption across key sectors such as education, health, energy, and transportation [1], [3]. Complementing this strategy, the government has established specialized institutions, including the École Nationale Supérieure d'Intelligence

* Corresponding author.
besma.ahmedmalek@univ-bba.dz (B. AHMED MALEK); ahmed-malek-nada@univ-eltarf.dz (N. AHMED MALEK)

Artificielle (ENSIA) and the École Nationale Supérieure de Mathématiques (ENSM), to cultivate a generation of AI experts and innovators [4], [5]. These efforts reflect a commitment to building a future-ready workforce capable of meeting the demands of a rapidly evolving global economy.

In addition to these educational advancements, sectoral initiatives have demonstrated the practical application of AI in workforce development. For instance, Algérie Télécom's ESABOURA platform offers tailored virtual learning solutions to upskill employees in telecommunications and other domains [6]. Similarly, Sonatrach's collaborations with global technology leaders like Huawei integrate AI into employee training programs, enhancing operational efficiency and technical competencies [7].

However, despite these advancements, Algeria faces persistent challenges. Limited digital infrastructure, particularly in rural areas, restricts access to AI-driven training platforms, exacerbating socio-economic disparities [1], [3]. Furthermore, policy fragmentation and a lack of stakeholder coordination impede the scalability and effectiveness of these programs. The absence of robust evaluation mechanisms to measure the impact of AI initiatives on workforce productivity further constrains the ability to refine and expand these efforts [8].

This paper aims to examine the role of AI in revolutionizing workforce development in Algeria, with a focus on employee training programs. By analyzing case studies from key industries, addressing systemic challenges, and proposing actionable recommendations, this study provides insights for policymakers, industry leaders, and educators to foster an inclusive and future-ready workforce.

## 2. Background and Literature Review

### 2.1. The Role of AI in Workforce Development

Artificial Intelligence (AI) is fundamentally reshaping workforce development by enabling organizations to address skill gaps, improve productivity, and prepare employees for dynamic workplace environments. Traditionally, workforce training programs have emphasized technical proficiency, often overlooking the critical role of soft skills. However, in an increasingly automated landscape, competencies such as adaptability, effective communication, and collaboration are becoming as essential as technical expertise. According to Hussain (2024), the integration of AI into workforce development accelerates the automation of routine tasks, shifting the focus to complex problem-solving and interpersonal abilities that machines cannot replicate [9].

### 2.1.1. Balancing Technical and Soft Skills

As routine tasks are increasingly automated, the ability to adapt to new roles, work in diverse teams, and solve complex problems is paramount. Tariq (2024) emphasizes that AI has revolutionized reskilling by enabling tailored training programs that address both technical and interpersonal competencies [5]. For instance, companies like Siemens have implemented AI-enabled training platforms that combine technical reskilling with programs to enhance interpersonal skills, resulting in improved performance and retention rates. Similarly, AI-driven tools such as virtual simulations provide employees with opportunities to practice real-world scenarios, such as conflict resolution or negotiation, fostering a balance between technical and interpersonal skills [9].

#### 2.1.2. Continuous Learning Through AI

AI facilitates lifelong learning by offering personalized, adaptive training experiences. Unlike traditional methods, AI-driven platforms dynamically assess employee progress, identify skill gaps, and deliver customized content. As Lokesh et al. (2024) argue, continuous learning is critical in the age of technological shifts, where the rapid evolution of AI demands constant updates to workforce skills [10]. Platforms such as Coursera and LinkedIn Learning employ AI to curate training modules based on individual career goals and industry trends, enabling employees to remain competitive in fast-changing markets. This model is particularly vital in Algeria, where aligning education with labor market needs remains a persistent challenge [5].

#### 2.1.3. Overcoming Barriers to AI-Driven Training

While AI presents significant opportunities for workforce development, challenges persist. Limited digital infrastructure, particularly in rural areas, hinders the widespread adoption of AI-driven training platforms. Additionally, concerns about algorithmic bias in AI systems raise questions about fairness and inclusivity. Tariq (2024) highlights that to mitigate these challenges, organizations must adopt transparent AI systems, invest in equitable access to technology, and address potential biases in training algorithms [5].

#### 2.1.4. Strategic Impact of AI in Workforce Development

By integrating AI into workforce training, organizations can achieve far-reaching benefits:

- Increased Productivity: A workforce equipped with both technical and soft skills is more agile and innovative, driving organizational success [9].
- Resilient Culture: Emphasizing continuous learning and interpersonal skills fosters a culture that adapts to technological and market disruptions [10].
- Economic Growth: On a national scale, AI-driven workforce development supports economic diversification and competitiveness, particularly in regions like Algeria, where industries are transitioning towards digital transformation [5].

A holistic approach to workforce development, combining AI-driven technical reskilling with the cultivation of soft skills, is essential for navigating the complexities of the future workplace. By addressing infrastructure challenges, fostering inclusivity, and prioritizing continuous learning, organizations can harness AI to build a dynamic, future-ready workforce. As Algeria continues to adopt AI in key sectors such as telecommunications and hydrocarbons, these strategies will play a critical role in shaping its economic and social trajectory.

### 2.2. Current Workforce Challenges in Algeria

Algeria faces significant challenges in workforce development that hinder its ability to meet the demands of an evolving labor market. Despite notable efforts to modernize education and training systems, structural and systemic barriers persist, affecting both employee training and overall economic growth. The following subsections detail these challenges, providing insights into key areas that require attention.

#### 2.2.1. Skills Mismatch and Education-to-Employment Gap

- Inadequate Alignment with Industry Needs: Many training programs in Algeria fail to align with the specific skill requirements of industries, particularly in high-growth areas like information and communication technology (ICT) and energy. This mismatch leaves graduates underprepared for the demands of the job market [6].
- High NEET Rate: Approximately 26.2% of youth aged 15–24 are classified as NEET (Not in Employment, Education, or Training), reflecting a significant gap between education systems and workforce entry points [7].
- **Slow Integration of Advanced Technologies**: Educational and vocational institutions often struggle to incorporate AI, machine learning, and other emerging technologies into their curricula, further widening the skills gap.

#### 2.2.2. Limited Digital Infrastructure

- Urban-Rural Divide: Digital infrastructure in rural areas remains underdeveloped, limiting access to AI-driven learning platforms and online training resources. This disparity exacerbates inequalities in workforce development [7].
- Low Internet Penetration: While urban centers benefit from moderate internet coverage, overall connectivity levels lag behind global standards, hindering the implementation of scalable digital training programs.

#### 2.2.3. Talent Drain and Retention Issues

- Brain Drain: Algeria experiences significant emigration of skilled professionals seeking better opportunities abroad, particularly in fields like ICT and engineering. This exodus leaves industries struggling to fill critical roles [8].
- Retention Challenges: Companies face difficulties retaining trained employees due to limited career advancement opportunities and competitive salaries in neighboring regions or international markets.

#### 2.2.4. Resistance to Technological Change

- Cultural Barriers: Many organizations and employees are hesitant to adopt new technologies, fearing job displacement or the complexity of transitioning to AI-driven workflows [7].
- Traditional Training Methods: Conventional training approaches are still predominant, with limited adoption of AI-enabled, adaptive learning systems that could enhance training efficiency.

Addressing these workforce challenges requires a multi-faceted approach that combines policy reforms, investments in digital infrastructure, and enhanced collaboration among stakeholders. By aligning training programs with industry needs, fostering inclusivity in digital education, and encouraging public-private partnerships, Algeria can better prepare its workforce for the demands of a rapidly evolving global economy.

### 2.3. Adoption of AI in Algeria:

Algeria has embarked on significant initiatives to integrate Artificial Intelligence (AI) into its national development strategy, focusing on positioning itself as a regional leader in AI innovation and competitiveness. These efforts are reflected in strategic frameworks, educational advancements, and industry collaborations.

### 2.3.1. National Strategies

In May 2024, the Ministry of Digitalization and the Knowledge Economy officially launched Algeria's National Artificial Intelligence Strategy. This ambitious roadmap aims to harness AI's transformative potential to drive innovation and sustainable development. The strategy is built on five key pillars [9]:

- Regulatory and Legal Framework: Developing clear and adaptive laws to ensure the ethical and responsible progression of AI while protecting citizens' rights and encouraging innovation.
- Capacity Building in Research and Education: Establishing specialized programs for researchers and students to foster talent and creativity in AI.
- Promotion of Innovation and Entrepreneurship: Encouraging startups and enterprises to adopt and develop AI-based solutions.
- Technological Infrastructure Development: Establishing robust digital infrastructures to support AI applications.
- International Cooperation: Partnering with global institutions and companies to exchange knowledge and technologies.

### 2.3.2. Innovative Educational and Institutional Initiatives

To address workforce challenges and align educational outcomes with labor market demands, Algeria has introduced several educational innovations:

- Online Training for New Recruits in Higher Education: Newly recruited university instructors undergo online training programs to ensure they are equipped with modern pedagogical and technological skills. These programs leverage scalable digital platforms to improve teaching quality and integrate technology into education systems [11].
- Creation of Specialized AI and Math Schools: The establishment of the École Nationale Supérieure d'Intelligence Artificielle (ENSIA) and the École Nationale Supérieure de Mathématiques (ENSM) represents a strategic move to address the skills gap in AI and advanced technologies. These institutions aim to cultivate a new generation of AI experts and innovators, aligning educational outcomes with labor market demands [3].

### 2.3.3. Industry Partnerships for Workforce Training

Industry partnerships play a vital role in fostering AI adoption and enhancing workforce capabilities:

- Sonatrach and Global Technology Leaders: Algeria's national energy company, Sonatrach, has collaborated with global technology companies like Huawei to integrate AI into workforce training programs. These partnerships enable the adoption of predictive maintenance systems, enhance operational efficiency, and improve employee technical skills [4].
- Telecommunications Sector: Algérie Télécom has implemented AI-driven platforms such as ESABOURA to optimize customer service, enhance employee training, and improve overall network management [12].

### 2.3.4. Key Industries Adopting AI

Several sectors in Algeria have begun leveraging AI to enhance efficiency and innovation:

- Telecommunications: Algérie Télécom has implemented AI-driven platforms such as ESABOURA, which optimize customer service and network management [5].
- Energy and Hydrocarbons: Sonatrach has partnered with global technology leaders to integrate AI into exploration, production, and predictive maintenance [10].
- Healthcare: AI is being utilized for medical diagnostics and electronic patient record management, improving healthcare delivery [12].
- Agriculture: Startups are developing AI-powered solutions to optimize crop yields and manage water resources efficiently [3].

### 2.3.5. Progress to Date

While Algeria has made notable strides in adopting AI, challenges remain that hinder its full implementation:

- Limited Digital Infrastructure: Rural areas lack adequate digital infrastructure, restricting access to AI-driven solutions and training platforms [9], [3].
- Policy Fragmentation: A lack of coordination among stakeholders slows the effective execution of AI strategies [10].
- Skills Gap: There is a growing need for workforce training and skill development in AI to meet market demands [13].

Despite these challenges, Algeria has demonstrated progress through initiatives such as the establishment of specialized AI institutions like the École Nationale Supérieure d'Intelligence Artificielle (ENSIA) and collaborations with international technology firms [5]. The development of AI-focused frameworks and increasing private-public partnerships highlight Algeria's commitment to fostering AI adoption.

## 3. Methodology

This study adopts a mixed-methods approach to examine the role of Artificial Intelligence (AI) in workforce development in Algeria. It combines secondary data analysis with case studies to identify trends, challenges, and opportunities for AI-driven workforce training. The methodology is structured into three key components: data collection, analytical framework, and clarification of adoption rates.

### 3.1. Data Collection

Data for this research is drawn from a diverse set of credible secondary sources and real-world examples to ensure a balance of quantitative and qualitative insights:

- Government Reports:
  - National Artificial Intelligence Strategy (2024): Published by the Ministry of Digitalization and Knowledge Economy, this report outlines Algeria's goals for AI adoption and workforce transformation [12].
  - Annual Reports by the Autorité de Régulation de la Poste et des Communications Électroniques (ARPCE): These provide insights into digital

infrastructure and technological readiness across Algeria [5].The National Artificial Intelligence Strategy (2024) published by the Ministry of Digitalization and Knowledge Economy, which outlines Algeria's goals for AI adoption and workforce transformation [12].

- Industry Publications:
  - Algérie Télécom's 2023 Annual Report: Details platforms like ESABOURA and LabLabee, with data on adoption rates and training program effectiveness [11], [3], [15].
  - Sonatrach's 2020 Digital Transformation Report: Highlights AI-driven initiatives, including predictive maintenance and workforce training programs [3].
- Academic and Regional Studies:
  - European Training Foundation's Country Fiche Algeria (2022): Provides workforce development data and the potential for technological integration in Algeria [4].
  - Comparative Benchmarks from Tunisia and Morocco: Used to estimate adoption rates in sectors with limited direct Algerian data [4].
- Case Studies:
  - Algérie Télécom: Focus on the outcomes of AI-driven platforms like ESABOURA, LabLabee, and ELEPHORM.
  - Sonatrach: Examination of AI applications in predictive maintenance and operational efficiency.
  - Healthcare and Agriculture: Analysis of emerging AI-driven tools for diagnostics and precision farming.

### 3.2. Analytical Framework

The study utilizes a mixed-methods analytical approach, combining quantitative and qualitative methodologies:

#### 3.2.1. Quantitative Analysis:

- Workforce Indicators: NEET (Not in Employment, Education, or Training) rates and training participation metrics were analyzed using data from the European Training Foundation and industry reports [4].
- Adoption Rates of AI: Data was derived from Algérie Télécom and Sonatrach reports, supplemented by regional benchmarks for comparable industries [11], [3], [3].

#### 3.2.2. Qualitative Analysis:

- Policy Analysis: Thematic review of strategic documents, including the National Research and Innovation Strategy on Artificial Intelligence (2020–2030), to understand policy directions [12].
- Stakeholder Perspectives: Interviews and testimonials from policymakers, industry leaders, and educators provided insights into challenges and perceptions related to AI integration.

### 3.3. Clarification of Adoption Rates

Adoption rates presented in this study are based on a combination of direct data and extrapolations to address data availability gaps:

- Telecommunications Sector (65%): Derived from Algérie Télécom's reports on ESABOURA and LabLabee, which indicate significant adoption of AI-driven workforce training platforms [11], [3].
- Energy Sector (50%): Based on Sonatrach's documented integration of AI into operations, aligned with regional adoption trends observed in North Africa [3].
- Healthcare (30%) and Agriculture (25%): Extrapolated from global AI adoption rates and regional benchmarks, supported by academic studies and ETF reports [4].

These estimates are explicitly acknowledged as approximations, providing a directional understanding of sectoral trends in the absence of proprietary data. Future research should incorporate primary data collection for greater precision.

## 4. Results

This section presents the findings of the study, focusing on the current status of AI adoption in Algerian workforce training, the barriers to AI integration, and the preliminary impacts of AI-driven programs like ESABOURA and LabLabee. The results are supported by data from key industries and visual representations to enhance understanding.

### 4.1. Current Status of AI in Algerian Workforce Training

AI adoption in workforce training in Algeria is uneven across sectors, with telecommunications and energy demonstrating higher adoption rates compared to healthcare and agriculture. These disparities highlight varying levels of readiness and investment.

**Table 1. AI Adoption Rates in Workforce Training by Sector**

| Sector | Adoption Rate (%) | Examples of AI Integration | Key Outcomes |
|---|---|---|---|
| Telecommunications | 65% | ESABOURA, LabLabee (Algérie Télécom) | Improved training efficiency |
| Energy (Hydrocarbons) | 50% | Sonatrach's predictive maintenance programs | Enhanced operational safety |
| Healthcare | 30% | AI diagnostics and e-health initiatives | Faster and more accurate diagnostics |
| Agriculture | 25% | AI-driven precision farming tools | Optimized resource usage |

The telecommunications sector leads in AI adoption, driven by initiatives like ESABOURA and LabLabee, which focus on enhancing employee skills through AI-based training platforms [11], [3].

The energy sector has achieved moderate adoption, with AI applications improving workforce safety and operational efficiency [3].

Healthcare and agriculture lag behind due to limited digital infrastructure and lower investment in AI-driven solutions [4].

### 4.2. Barriers to AI Integration

Several barriers hinder the widespread adoption of AI in workforce training programs across Algeria.

**Table 2. Key Barriers to AI Integration**

| Category | Specific Challenges | Impact |
|---|---|---|
| Infrastructure | Limited internet access in rural areas | Reduced accessibility to AI platforms |
| Policy Fragmentation | Lack of coordinated AI policies | Slower implementation of strategies |
| Skills Gap | Insufficient technical and AI knowledge | Delayed workforce readiness |
| Cultural Resistance | Hesitancy to adopt AI-driven training | Slower organizational adaptation |

Infrastructure gaps, particularly in rural areas, restrict access to platforms like ESABOURA, limiting their reach and effectiveness [5].

Policy fragmentation and a lack of public-private coordination slow the scaling of AI initiatives [12].

The skills gap in technical and AI-related competencies remains a significant obstacle to workforce readiness [4].

### 4.3. Impact Assessment

AI-enabled training programs have demonstrated tangible benefits in improving workforce skills, efficiency, and adaptability.

**Table 3. Impact of AI-Enabled Programs**

| Program | Sector | Key Benefits | Challenges Faced |
|---|---|---|---|
| ESABOURA (Algérie Télécom) | Telecommunications | Increased training efficiency by 40% | Limited rural accessibility |
| LabLabee (Algérie Télécom) | Telecommunications | Enhanced technical skill development | High initial setup costs |
| Sonatrach AI Program | Energy | Reduced equipment downtime by 20% | Complex implementation requirements |
| AI Diagnostics | Healthcare | Improved diagnostic accuracy by 30% | Data privacy concerns |

ESABOURA and LabLabee have significantly increased training efficiency and skill development in the telecommunications sector, demonstrating the potential of AI to enhance workforce capabilities [11], [3].

Sonatrach's AI-driven maintenance programs have improved operational safety and reduced downtime, though implementation complexity remains a challenge [3].

Emerging healthcare AI tools show promise in improving diagnostic accuracy, but concerns about data privacy and infrastructure remain prevalent [4].

## 5. Discussion

The findings of this study reveal both significant progress and persistent challenges in the adoption of Artificial Intelligence (AI) for workforce development in Algeria. While sectors like telecommunications and energy have demonstrated considerable advancements, others, such as healthcare and agriculture, lag behind due to systemic barriers. This section discusses these findings in light of ethical considerations, infrastructural challenges, and the need for targeted strategies to maximize AI's potential across all sectors.

### 5.1. Addressing the Skills Gap

The results demonstrate a significant skills gap in AI-related competencies, particularly in sectors like healthcare and agriculture. While platforms like ESABOURA and LabLabee have proven effective in the telecommunications sector, their reach remains limited by disparities in access and infrastructure. Bridging this gap requires:

- National Education Reforms: Incorporating AI and digital literacy into school curricula to build foundational skills early.
- Sector-Specific Training Programs: Expanding AI-driven training initiatives beyond telecommunications and energy to sectors like healthcare and agriculture.
- Public-Private Partnerships: Collaborating with private enterprises to design and fund scalable training solutions tailored to industry needs.

These initiatives should prioritize inclusivity by addressing barriers for marginalized groups, particularly women, rural populations, and low-income communities.

### 5.2. Overcoming Infrastructure Barriers

Infrastructure remains a critical bottleneck for AI adoption, particularly in rural areas where internet access is limited. This disparity restricts the potential impact of AI-driven platforms like ESABOURA. To address this challenge:

- Investing in Digital Infrastructure: The government must prioritize rural broadband expansion to enable equitable access to AI platforms.
- Mobile-First AI Solutions: Developing lightweight, mobile-friendly platforms to reach underserved regions with limited connectivity.
- Incentivizing Private Sector Involvement: Offering tax breaks or subsidies to telecommunications providers for expanding infrastructure to remote areas.

Infrastructure development should align with Algeria’s National Artificial Intelligence Strategy (2024) to ensure coherence and scalability.

### 5.3. Acknowledging Data Limitations

This study relies on extrapolated adoption rates for healthcare (30%) and agriculture (25%) due to limited proprietary data. While these estimates provide a directional understanding, they may not fully capture the nuances of AI adoption in Algeria. Future research should incorporate primary data collection through interviews or surveys to refine these estimates and validate findings.

### 5.4. Enhancing Policy and Coordination

Policy fragmentation and a lack of coordination among stakeholders emerged as key challenges hindering the scalability of AI-driven workforce initiatives. Addressing these issues requires:

- Unified National Strategies: Streamlining policies and aligning them with the National Artificial Intelligence Strategy (2024) to foster consistency and coherence.

- Establishing an AI Task Force: Creating a multidisciplinary task force to oversee and coordinate AI adoption across sectors.
- International Collaboration: Partnering with global AI leaders to adopt best practices and integrate proven models for workforce development.

This task force should also monitor compliance with ethical standards, particularly in data handling and inclusivity.

### 5.5. Scaling AI Impact Across Sectors

While telecommunications and energy sectors have demonstrated moderate to high levels of AI adoption, the results reveal significant untapped potential in healthcare and agriculture. Key strategies to scale AI adoption include:

- Healthcare: Expanding the use of AI-driven diagnostic tools and telemedicine platforms to improve healthcare access and accuracy in rural areas.
- Agriculture: Promoting precision farming technologies to optimize resource utilization and increase productivity in the agricultural sector.

These targeted approaches can diversify Algeria's workforce capabilities while fostering sustainable economic growth.

### 5.6. Ethical and Cultural Considerations

AI adoption must address ethical and cultural challenges to ensure long-term success and societal acceptance:

- Data Privacy: Current frameworks are insufficient for protecting sensitive workforce and healthcare data. The lack of robust legal protections increases the risk of misuse or breaches [12].
    - Actionable Strategy: Introduce comprehensive data protection laws aligned with international standards to ensure ethical AI usage.
- Fairness and Inclusivity: AI initiatives must address disparities by designing accessible platforms for marginalized groups, including women, rural populations, and low-income communities [4].
    - Actionable Strategy: Develop AI training programs that accommodate diverse linguistic, socio-economic, and technological needs.
- Transparency and Accountability: Algorithms used in workforce training platforms should be auditable and transparent to avoid biases that could perpetuate inequality.
    - Actionable Strategy: Implement mandatory bias audits and transparency standards for AI systems.

Ethical considerations must be embedded into all AI policies to ensure that adoption is equitable and sustainable.

### 5.7. Strategic Recommendations

Based on the results and discussion, the following strategies are proposed to enhance AI adoption in workforce development:

- Develop a National AI Education Framework: Incorporate AI literacy into all levels of education and vocational training programs.

- Expand Public-Private Partnerships: Foster collaborations between government, academia, and the private sector to co-develop scalable and inclusive AI solutions.
- Invest in Research and Innovation: Establish AI research hubs to drive innovation and tailor AI applications to Algeria's unique workforce challenges.
- Monitor and Evaluate AI Programs: Implement robust evaluation mechanisms to assess the effectiveness and scalability of AI-driven workforce training programs.

## 6. Conclusion

The integration of Artificial Intelligence (AI) into workforce development represents a pivotal opportunity for Algeria to modernize its labor market, address systemic inefficiencies, and align with global technological advancements. This study highlights the transformative potential of AI in key sectors such as telecommunications and energy, driven by initiatives like Algérie Télécom's ESABOURA and LabLabee platforms, and Sonatrach's predictive maintenance programs. However, it also uncovers significant barriers, including skills gaps, limited infrastructure, policy fragmentation, and cultural resistance, which hinder the scalability and effectiveness of these initiatives.

The findings underscore the need for a holistic approach to AI adoption that goes beyond technical training to include:

- Comprehensive education reforms to build AI literacy and align academic programs with industry demands.
- Strategic investments in digital infrastructure, particularly in underserved rural areas, to ensure equitable access to AI-driven training platforms.
- Policy coherence and enhanced coordination among stakeholders to streamline AI implementation across sectors.

Ethical considerations, such as data privacy and inclusivity, to foster societal trust and acceptance of AI technologies.

By addressing these challenges and fostering an ecosystem that integrates public-private partnerships, research innovation, and international collaboration, Algeria can unlock the full potential of AI to revolutionize workforce development. This approach will not only enhance individual and organizational productivity but also support Algeria's broader goals of economic diversification and sustainable growth.

Future research should focus on longitudinal studies to assess the long-term impacts of AI on workforce development and explore innovative applications of AI in emerging sectors such as healthcare and agriculture. By leveraging these insights, Algeria can position itself as a regional leader in AI-driven workforce transformation, setting a benchmark for other nations in the region.

## References

[1] Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, "Stratégie Nationale de Recherche et d'Innovation en Intelligence Artificielle 2020–2030," Alger, Algérie, 2020. [Online]. Available: https://www.aps.dz/sante-science-technologie/116102.

[2] Algérie Presse Service (APS), "Écoles supérieures des mathématiques et de l'intelligence artificielle : une formation d'élite aux normes universelles," 2022. [Online]. Available: https://www.aps.dz/sante-science-technologie/127581.

[3] Algérie Télécom, "ESABOURA Platform Overview," Annual Report, 2023. [Online]. Available: https://www.algerietelecom.dz.

[4] Sonatrach, “Annual Report 2020: Digital Transformation in the Oil and Gas Industry,” Alger, Algérie, 2020. [Online]. Available: https://www.sonatrach.com.

[5] European Training Foundation, "Country Fiche Algeria 2022," 2022. [Online]. Available: https://www.etf.europa.eu..

[6] M.A. Hussain, “The Impact of Artificial Intelligence on Workforce Automation and Skill Development,” *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 2024.

[7] M.U. Tariq, “The Role of AI in Skilling, Upskilling, and Reskilling the Workforce,” 2024.

[8] G.R. Lokesh, K.S. Harish, V.S. Sangu, S. Prabakar, V.S. Kumar, and M. Vallabhaneni, “AI and the Future of Work: Preparing the Workforce for Technological Shifts and Skill Evolution,” 2024.

[9] M.A. Hussain, “The Impact of Artificial Intelligence on Workforce Automation and Skill Development,” *Journal of Artificial Intelligence, Machine Learning and Neural Network*, 2024.

[10] M.U. Tariq, “The Role of AI in Skilling, Upskilling, and Reskilling the Workforce,” 2024.

[11] Ministère de la Numérisation et de l’Économie du Savoir, “Stratégie nationale pour l’intelligence artificielle,” Alger, Algérie, 2024.

[12] Algérie Presse Service (APS), “AI in Healthcare: Transforming Patient Care,” 2023. [Online]. Available: https://www.aps.dz.

[13] Ministère de la Numérisation et de l'Économie du Savoir, “Stratégie nationale pour l’intelligence artificielle,” Alger, Algérie, 2024. [Online]. Available: https://www.ministere-numerisation.gov.dz/.

[14] Autorité de Régulation de la Poste et des Communications Électroniques (ARPCE), “Rapport annuel 2022 : État des infrastructures numériques,” Alger, Algérie, 2022. [Online]. Available: https://www.arpce.dz.

[15] Algérie Télécom, “LabLabee Platform Overview,” Annual Report, 2023. [Online]. Available: https://www.algerietelecom.dz.

# SE-GNN: A Social-Enhanced Graph Neural Network for Personalized Recommendation

SaraGasmi[1,†],SafaGasmi[,*,†]and TaharBouhadada[3,†]

[1]Badji Mokhtar University, LRI Laboratory, Computer Science Department, Annaba 23000, Algeria
[2] Badji Mokhtar University, LRI Laboratory, Computer Science Department, Annaba 23000, Algeria
[3] Badji Mokhtar University, LRI Laboratory, Computer Science Department, Annaba 23000, Algeria

**Abstract**

Social recommendation has emerged as a promising approach to enhancing the accuracy and relevance of personalized suggestions by leveraging users' social connections and influences. However, effectively modeling the complex relationships between users, items, and social ties remains a key challenge. In this work, we propose SE-GNN (Social-Enhanced Graph Neural Network), a novel graph-based recommendation model that seamlessly integrates social information and user-item interactions within a unified framework. SE-GNN's sophisticated dual-channel architecture dynamically fuses social influence patterns and user preference signals to provide more personalized recommendations. Extensive experiments on the Epinions and Ciao datasets demonstrate the superior performance of SE-GNN compared to state-of-the-art social recommendation methods. An in-depth ablation study further highlights the importance of the model's key components in achieving these improvements.

**Keywords**

Recommender systems, social recommendation, graph neural networks, dual-channel architecture, adaptive fusion

## 1. Introduction

Recommendation systems have become an integral part of our daily digital experiences, helping us navigate the overwhelming abundance of content and products available online. These systems aim to provide personalized suggestions that cater to individual users' preferences and interests. However, traditional recommendation approaches often overlook the valuable social context that exists among users, which can significantly influence their preferences and behaviors [1]. The growing prominence of social media platforms has underscored the importance of incorporating social information into recommendation systems. Users' connections, interactions, and social influences can provide valuable insights that can enhance the accuracy and relevance of recommended items [2, 3]. Recognizing this, researchers have actively explored ways to leverage social data to improve the performance of recommendation systems, leading to the emergence of the field of social recommender systems [4, 5]. One of the key challenges in social recommendation is effectively modeling and integrating the complex relationships between users, items, and social connections. Conventional methods have often relied on matrix factorization or graph-based approaches, which may struggle to capture the nuanced interplay between these diverse entities[6].

In this work, we propose SE-GNN (Social-Enhanced Graph Neural Network), a novel approach that leverages the power of graph neural networks to jointly model user-item interactions and social relationships within a unified framework. By designing a sophisticated dual-channel architecture that adaptively combines social influence patterns and user preference signals, SE-GNN aims to overcome the limitations of previous social recommendation methods and provide more accurate and personalized recommendations [7, 8].

The main contributions of this paper are as follows:

1. We introduce SE-GNN, a novel social recommendation model that seamlessly integrates social information and user-item interactions through a dual-channel message passing mechanism.

2. We develop an adaptive fusion strategy that allows the model to dynamically balance the relative importance of social influence and personal preferences for different users.

3. We conduct extensive experiments on two widely-used social recommendation datasets, Epinions and Ciao, and demonstrate the superior performance of SE-GNN compared to state-of-the-art social recommendation methods.

4. We perform an in-depth ablation study to evaluate the individual contributions of the key components of SE-GNN, providing insights into the model's design and effectiveness.

The remainder of the paper is organized as follows. Section 2 reviews the related work on social recommendation and graph neural networks. Section 3 presents the detailed architecture and training of the proposed SE-GNN model. Section 4 describes the experimental setup, datasets, and evaluation metrics, followed by the results and analysis in Section 5. Finally, Section 6 concludes the paper and discusses potential future research directions.

## 2. Background

This section reviews key related works, focusing on recommender systems, graph neural networks (GNN)-based recommendation methods

### 2.1. Recommendation system

Recommender systems have emerged as essential information filtering tools in the digital age, designed to address the information overload challenge by providing personalized suggestions to users. These systems analyze vast amounts of data to predict user preferences and recommend relevant items (products, services, or content) by leveraging various information sources: historical user interactions, item characteristics, and contextual information [9]. Through sophisticated algorithms, they establish meaningful connections between users and items, either by identifying patterns in user behavior (collaborative filtering), analyzing item features (content-based filtering), or combining multiple approaches (hybrid systems) [10]. The primary objective is to enhance user experience by reducing search efforts and providing serendipitous discoveries while simultaneously achieving business goals such as increased user engagement and satisfaction [11]. In essence, recommender systems serve as automated decision support systems that continuously learn and adapt to evolving user preferences to deliver increasingly accurate and personalized recommendations.

### 2.2. Graph neural network

A Graph Neural Network (GNN) is a class of deep learning models specifically designed to process data represented as graphs [12]. Unlike traditional neural networks that operate on grid-structured data, GNNs are capable of leveraging the non-Euclidean structure of graphs, where nodes represent entities and edges capture relationships between them. By iteratively aggregating and transforming features from a node’s local neighborhood, GNNs enable the representation and learning of rich relational information within the graph [13].

The core architecture of GNNs typically includes message passing mechanisms, where nodes exchange information with their neighbors to update their feature representations, followed by readout functions to generate outputs for tasks such as node classification, graph classification, or link prediction. These models have demonstrated superior performance in a wide range of applications, including social network analysis, recommendation systems, molecular property prediction, and knowledge graph reasoning. GNNs are an active area of research, with advancements focusing on improving scalability, interpretability, and generalization to diverse graph structures, highlighting their critical role in advancing machine learning for structured data.

### 2.3. Graph neural networks for recommendation systems

Recent advances in deep learning have revolutionized the field of recommendation systems, with Graph Neural Networks (GNNs) emerging as a particularly powerful paradigm. GNNs have demonstrated remarkable capabilities in capturing complex user-item interactions and leveraging the inherent graph structure of recommendation data [14]. This section explores the fundamental reasons behind the effectiveness of GNNs in recommendation systems and their key advantages over traditional approaches.

#### 2.3.1. Natural Representation of User-Item Interactions

Recommendation systems inherently deal with user-item interactions that can be naturally modeled as bipartite graphs, where users and items represent nodes, and interactions form edges. GNNs excel in processing such structured data by leveraging the graph topology to learn meaningful representations. Unlike traditional matrix factorization methods, GNNs can capture higher-order connectivity patterns through message passing between nodes, enabling a more comprehensive understanding of user preferences and item relationships [15].

#### 2.3.2. Enhanced Feature Propagation and Aggregation

The message-passing mechanism in GNNs facilitates efficient information flow across the user-item graph. Through multiple layers of neighborhood aggregation, GNNs can effectively capture both local and global interaction patterns. This hierarchical feature learning process enables the model to:

Propagate user preferences through the item-user-item paths
Aggregate similar user behaviors to enhance recommendation accuracy
Learn latent features that represent both explicit and implicit relationships

#### 2.3.3. Integration of Heterogeneous Information

Modern recommendation scenarios often involve diverse types of information beyond simple user-item interactions. GNNs provide a flexible framework for incorporating:

- Multiple types of interactions (purchases, views, ratings)
- Content features of items and user profiles
- Temporal dynamics of user behavior
- Social relationships between users

This ability to handle heterogeneous information sources in a unified framework has led to significant improvements in recommendation quality [16].

#### 2.3.4. Performance Advantages

Empirical studies have demonstrated several key advantages of GNN-based recommendation systems:

1. **Higher Accuracy**: GNNs consistently outperform traditional methods in terms of ranking metrics (NDCG, Precision@K)

2. **Better Cold-Start Handling**: The graph structure helps in making recommendations for new users/items
3. **Improved Interpretability**: The message-passing mechanism provides insights into recommendation logic
4. **Scalability**: Modern GNN architectures can efficiently handle large-scale recommendation scenarios

### 2.3.5. Key Applications and Impact

The effectiveness of GNNs in recommendation systems has been validated in various real-world applications:

- E-commerce platforms: Product recommendation and user behavior prediction
- Social networks: Friend suggestion and content recommendation
- Streaming services: Media content recommendation
- Online advertising: Click-through rate prediction

These applications have demonstrated significant improvements in key business metrics, including user engagement, conversion rates, and customer satisfaction [17].

## 3. SE-GNN: A Novel Social-Enhanced Graph Neural Network for Recommendation Systems

We propose SE-GNN (Social-Enhanced Graph Neural Network), a novel approach that leverages the power of graph neural networks to capture both explicit social relationships and implicit user-item interactions within a unified framework. While conventional recommendation systems often struggle to effectively integrate social information with user behavior patterns, SE-GNN addresses this limitation through an innovative architectural design that seamlessly combines these crucial information sources. Our model addresses key challenges in social recommendation by introducing a sophisticated dual-channel architecture that effectively processes and combines social influence patterns with user preference signals, thereby significantly improving recommendation accuracy and user engagement.

The foundation of SE-GNN lies in its heterogeneous graph structure G = (V, E), which encompasses both user-item interactions and social connections in a comprehensive manner. This graph-based representation serves as a powerful abstraction of the complex social recommendation space, enabling our model to capture and process multi-dimensional relationships simultaneously. The vertex set V comprises users U = {$u_1$, $u_2$, ..., $u_n$} and items I = {$i_1$, $i_2$, ..., $i_m$}, representing the primary entities in our recommendation ecosystem. Here, each user u_i encapsulates individual preferences, behavioral patterns, and social network positions, while each item i_j represents products, content, or services with their associated features and interaction histories.

The edge set E consists of two distinct but interconnected components: user-item interaction edges E_ui and social connection edges E_s. The user-item interaction edges E_ui capture various forms of user engagement with items, including explicit feedback (such as ratings, reviews, and purchases) and implicit feedback (such as clicks, views, and browsing time). These interactions are crucial for understanding individual user preferences and consumption patterns. Simultaneously, the social connection edges E_s represent the complex web of social relationships between users, including direct connections (such as friendships and followings) and indirect relationships (such as membership in common communities or similar interest groups).

This sophisticated graph representation allows our model to capture complex relationships that exist in real-world recommendation scenarios in several innovative ways[14]. At the core of SE-GNN is an innovative dual-channel message passing mechanism that processes information through distinct yet complementary pathways:

1. Social Channel Aggregation: The social channel aggregates information from users' social connections through a sophisticated attention mechanism:

$$h\_u\text{\^{}}(s) = AGG\_s(\{TRANSFORM\_s(h\_v) \cdot \alpha\_uv \mid v \in N\_s(u)\}) \tag{1}$$

This formula represents the aggregation of transformed feature vectors from social neighbors, weighted by attention coefficients.
The TRANSFORM_s function applies a learnable transformation to neighbor features, while AGG_s combines the weighted features through a permutation-invariant aggregator.

2. Social Attention Weights: The attention weights between users are computed as:

$$\alpha\_uv = \text{softmax}(W\_a[h\_u \,||\, h\_v] + b\_a) \tag{2}$$

This attention mechanism learns to assign different importance weights to different social connections based on user embeddings. The concatenation operation [h_u || h_v] captures the pairwise relationship between users.

3. Interaction Channel Processing: The interaction channel processes user-item relationships through:

$$h\_u\text{\^{}}(i) = AGG\_i(\{TRANSFORM\_i(h\_v) \cdot \beta\_ui \mid v \in N\_i(u)\}) \tag{3}$$

This formula aggregates information from user-item interactions, employing a separate transformation and attention mechanism specific to interaction patterns.

4. Interaction Attention Weights: The interaction strength between users and items is captured by:

$$\beta\_ui = \text{softmax}(W\_b[h\_u \,||\, h\_i] + b\_b) \tag{4}$$

This attention mechanism learns to weight the importance of different user-item interactions based on their feature compatibility.

5. Adaptive Fusion Mechanism: The model combines social and interaction signals through:

$$h\_u = g\_u \odot h\_u\text{\^{}}(s) + (1 - g\_u) \odot h\_u\text{\^{}}(i) \tag{5}$$

$$g\_u = \text{sigmoid}(W\_g[h\_u\text{\^{}}(s) \,||\, h\_u\text{\^{}}(i)] + b\_g) \tag{6}$$

This adaptive fusion mechanism learns to balance the influence of social and interaction signals dynamically for each user. The gating vector g_u determines the relative importance of each channel.

6. Multi-objective Loss Function: The model optimization employs a comprehensive loss function:

$$L = L\_rec + \lambda_1 L\_social + \lambda_2 L\_reg \tag{7}$$

Where:

$$L\_rec = -\sum(u,i,j)\in D \log(\sigma(\hat{y}\_ui - \hat{y}\_uj)) \tag{8}$$

$$L\_social = \sum(u,v)\in E\_s \,||h\_u - h\_v||^2 \tag{9}$$

$$L\_reg = ||\Theta||^2 \tag{10}$$

The loss function combines recommendation accuracy (L_rec), social consistency (L_social), and regularization (L_reg) terms. L_rec implements Bayesian Personalized Ranking loss, L_social enforces similarity between socially connected users, and L_reg prevents overfitting.

7. Final Prediction: The prediction layer generates recommendation scores through:

$$\hat{y}\_ui = MLP([h\_u \,||\, h\_i]) \tag{11}$$

This final layer captures non-linear interactions between user and item representations through a Multi-Layer Perceptron (MLP), producing the final recommendation scores.
Each component of SE-GNN is carefully designed to capture different aspects of the recommendation problem, from social influence to user-item interactions, while maintaining computational efficiency and model expressiveness.

## 4. Experimental Evaluation of SE-GNN on Epinions and Ciao Datasets

To validate the effectiveness of the SE-GNN (Social-Enhanced Graph Neural Network) approach, we conducted experiments on two widely-used social recommendation datasets: Epinions and Ciao.
Datasets:
We utilized two benchmark datasets for our experiments, summarized in Table 1:

**Tabel 1** benchmark datasets for our experiments

| Dataset | Users | Items | Trust Relationships |
|---|---|---|---|
| Epinions | 40,163 | 139,738 | 664,824 |
| Ciao | 19,017 | 12,375 | 223,493 |

- Epinions: Contains user-item ratings and social trust relationships among users.
- Ciao: Similar to Epinions, it includes user-item ratings and social connections.

These datasets represent real-world social recommendation scenarios, making them ideal for evaluating the performance of SE-GNN.

Experimental Setup:
Our experimental methodology was as follows:

1. Dataset Splitting: Each dataset was divided into training, validation, and test sets using an 80/10/10 ratio.
2. Baselines and Implementation: SE-GNN was implemented alongside state-of-the-art social recommendation baselines, including SoRec, SoReg, and SBPR.
3. Model Parameters:

- User and item embedding dimensionality: 64.
- Prediction layer: A 2-layer MLP.
- Optimization: BPR (Bayesian Personalized Ranking) loss with social consistency regularization.

4. Hyperparameter Tuning: Parameters such as learning rate, regularization coefficients, and attention mechanism weights were tuned using the validation set.

Evaluation Metrics:
The performance of SE-GNN and baselines was assessed using the following metrics:
Recall@k: Measures the proportion of relevant items in the top-k recommended list.

$$\text{Recall@k} = \frac{\text{Number of relevant items in the top}-\text{k recommendations}}{\text{Total number of relevant items for the user}} \tag{12}$$

In our experiments, we set k=10, indicating a focus on the top-10 recommendations.
Normalized Discounted Cumulative Gain (NDCG@k): NDCG@k evaluates the ranking quality of recommended items, prioritizing the placement of relevant items higher in the ranked list. The discounted cumulative gain (DCG@k) is given by:

$$\mathrm{DCG@k}=\sum_{i=1}^{k}\frac{\mathrm{reli}}{\log 2(i+1)} \qquad (13)$$

Where reli is the relevance of the i-th item. NDCG@k normalizes DCG@k by the ideal DCG (IDCG@k).

.

## 5. Results

The results of SE-GNN on the Epinions and Ciao datasets, compared to the best-performing baseline SBPR, are summarized in **Table 2**:

**Table 2** the best-performing baseline SBPR

| Dataset | Metric | SE-GNN | SBPR (Best Baseline) |
|---|---|---|---|
| Epinions | Recall@10 | 0.3421 | 0.2984 |
| | NDCG@10 | 0.3987 | 0.3655 |
| Ciao | Recall@10 | 0.2845 | 0.2534 |
| | NDCG@10 | 0.3452 | 0.3128 |

The results demonstrate that SE-GNN effectively leverages both social connections and user-item interactions to provide significantly more accurate and relevant recommendations compared to state-of-the-art social recommendation methods.
Ablation Study:

To assess the contributions of the key components of SE-GNN, we conducted an ablation study with the following configurations:

- Social Channel Only: Utilizes only the social connections for recommendations.
- Interaction Channel Only: Considers only user-item interactions.
- Full SE-GNN Model: Combines both channels using an adaptive fusion mechanism.

The ablation results are presented in Table 3:

**Table 3** the ablation results

| Model Configuration | Recall@10 | NDCG@10 |
|---|---|---|
| Social Channel Only | 0.3163 | 0.3701 |
| Interaction Channel Only | 0.3274 | 0.3821 |
| Full SE-GNN Model | 0.3421 | 0.3987 |

The results show that the dual-channel architecture and the adaptive fusion mechanism are essential for SE-GNN's superior performance, as they allow the model to effectively balance social influence and user preferences.

In conclusion, the experimental evaluation on the Epinions and Ciao datasets demonstrates the effectiveness of the proposed SE-GNN approach in leveraging social information to enhance recommendation performance compared to state-of-the-art baselines.

## 6. Conclusion

In this work, we have proposed SE-GNN, a novel social recommendation model that effectively leverages both social connections and user-item interactions to provide personalized and accurate recommendations. The key innovations of SE-GNN include its dual-channel architecture that separately processes social influence and user preference signals, and the adaptive fusion mechanism that dynamically balances these two components.

Through extensive experiments on the Epinions and Ciao datasets, we have demonstrated the superior performance of SE-GNN compared to state-of-the-art social recommendation methods. The ablation study has further highlighted the individual contributions of the model's core components, providing insights into the design choices that enable SE-GNN to outperform previous approaches.

Going forward, there are several promising research directions to explore. Incorporating additional types of social information, such as user sentiments and content-based features, could potentially enhance the model's understanding of social influence. Exploring dynamic, time-aware social relationships may also lead to improved recommendation accuracy. Additionally, extending the SE-GNN framework to handle other types of relational data, such as knowledge graphs or heterogeneous networks, could broaden its applicability to a wider range of recommendation and prediction tasks.

Overall, the proposed SE-GNN model represents a significant advancement in social recommendation, showcasing the power of graph neural networks in effectively modeling the complex interplay between social and preference signals. We believe this work will inspire further research into developing sophisticated, socially-aware recommendation systems that can deliver increasingly personalized and relevant experiences for users.

## References

[1] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Techniques, applications, and challenges," *Recommender systems handbook,* pp. 1-35, 2021.

[2] J. Tang, X. Hu, and H. Liu, "Social recommendation: a review," *Social Network Analysis and Mining,* vol. 3, pp. 1113-1133, 2013.

[3] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE international conference on data mining*, 2008, pp. 263-272.

[4] H. Ma, H. Yang, M. R. Lyu, and I. King, "Sorec: social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 931-940.

[5] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang, "Social contextual recommendation," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 45-54.

[6] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer,* vol. 42, pp. 30-37, 2009.

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907,* 2016.

[8] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems,* vol. 30, 2017.

[9] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*, ed: Springer, 2010, pp. 1-35.

[10] R. Burke, "Hybrid web recommender systems," *The adaptive web: methods and strategies of web personalization,* pp. 377-408, 2007.

[11] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering,* vol. 17, pp. 734-749, 2005.

[12] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia, "Graph networks as learnable physics engines for inference and control," in *International conference on machine learning*, 2018, pp. 4470-4479.

[13] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems,* vol. 29, 2016.

[14] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 2019, pp. 165-174.

[15] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020, pp. 639-648.

[16] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, and T. Tan, "Session-based recommendation with graph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 346-353.

[17] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 974-983.

.

# VerChain: Blockchain Based Certificate Degree Attestation and Verification in Algeria

Rofaida Khemaissia[1,*,†] and Ala Djeddai[2,†]

[1] *Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa,12002, Algeria*

[2] *Laboratory of Computer Science and Applied Mathematics (LCSAM), Chadli Bendjedid El-Tarf University, B.P 73, El Tarf 36000, Algeria*

## Abstract

Many companies from all over the world are prone to counterfeit academic certificates, which would trigger colossal material losses, and it is time-consuming for universities to undertake this process by spending a huge number of budgets annually. In this case, the need for a degree verification process becomes important, as a famous platform Block-cert issued by Massachusetts Institute of Technology, aims to facilitate degree verification between the university and the company, empowered with the adoption of blockchain technology as a proof of existence by leveraging the immutability and availability features. In the academic and industrial community, several degree verification platforms have been proposed and their feasibility proved; however, in this paper, we propose the VerChain Algerian degree verification platform as a vital step towards fighting degree forgery and digitalization, which calls for the interference of several government ministries to render the process realistic and practical. VerChain is a blockchain based system for certified degree attestation and verification that aims to maintain security and data privacy, where it is managed by several Algerian ministries in which everyone has its responsibilities and access restrictions.

## Keywords

Blockchain, Hyperledger Fabric, Data Integrity, Smart Contracts, Privacy, Certificate attestation and verification.

## 1. Introduction

Under the supervision of the Algerian Minister of High Education and Scientific Research, 377,000 Algerian students graduate annually from public or private universities. Due to this huge number, enormous budgets are spent by the Algerian government to verify the validity and existence of a student's degree, let alone the wasted time beyond the process. Digitalization economizes costs and renders the process more trustworthy and rapid.

Blockchain is a well-known technology that has become widespread because of several concerns related to a trustless environment. According to the level of permission blockchain can be seen as two main categories, permissionless and permissioned, regarding the access rights to the network and the level of centralization, permissionless supports a wide public network in which every participant can partake and join the network as a blockchain peer and acts as a miner. Otherwise, a permissioned blockchain is known as private ledger, in which access to the network is under control and restricted only to legal nodes; it is known as semi-centralized.

[*] Corresponding author.

[†] These authors contributed equally.

🖂 khemaissia.rofaida@univ-tebessa.dz (R. Khemaissia) ;a.djeddai@univ-eltarf.dz (A. Djeddai) ;

iD 0000-0002-4092-537X (R. Khemaissia); 0000-0002-9354-4108 (A. Djeddai)

By leveraging different blockchain ‘merits such as immutability that helps to remain unalterable information, besides to auditability, availability and persistency. By adopting this technology, numerous academic and industrial solutions have been introduced for lucrative and free use. As centralizing the verification process is not supported by the Algerian government, the decentralization feature of the distributed ledger blockchain can handle a bulk of issues.

The research introduces VerChain, an Algerian framework for verifying degree certificates. This solution implements a system of separated roles to minimize potential errors and enhance both security and privacy measures. VerChain consists of three main actors starting with the ministries; then, the certificate issuer (i.e., University) and certificate verifier (national company), all of which are managed by blockchain, which acts as a mediator between actors. Because Blockchain maintains lightweight information for future scalability issues, VerChain replaced the certificate degree hard copy with a one-way generated hash as a proof of existence (SHA 256 algorithm). The hash code is generated by mashing several pieces of information to garner a particular hash that represents a unique identifier (ID) of a given certificate, thereby preserving the degree integrity and privacy of the student credentials. Blockchain technology is adopted as a distributed and decentralized database; however, altering (write, update, or delete) with saved information is more likely impossible because the information is replicated across a peer-to-peer network, which would help to increase availability and accessibility at any request. Since the National Authority for the Protection of Personal Data (ANPDP) [1] imposes that whatever the operators must undergo the privacy rules, students' credentials are provided privately by the Ministry of the Interior, Local Authorities, and Regional Planning, and the certificate verifier must undergo the authority rules by using the required information that the law allows.

The rest of the paper covers related works in Section 2, where Section 3 provides details about VerChain architecture and its main components, and Section 3 presents an example of the VerChain scenario executed using smart contracts. A possible implementation is presented in Section 4. Finally, Section 5 summarizes the paper.

## 2. Related Works

Digitalization is considered one step ahead for quick and better information processing, and it prevails in different fields, encompassing distance learning that grants students the opportunity to garner an online diploma. From here, the rate of degree or diploma falsification is increasing, that would be the reason behind adopting blockchain to create a trustworthy environment by many research, since The Massachusetts Institute of Technology (MIT) has put forward BlockCerts [2] an open and freely standard for academic certificate verification that was implemented on bitcoin as a public ledger, the system boosts a decentralized and trust student credentials verification via sending invitations to whom willing to join the system, through creating a student accounts then share it with the corresponding university/institute, where issuing only the degree onto the BlockCerts, in order to keep the credentials integrity BlockCerts through substituting the genuine degree hard copy with a generated irreversible hash code by employing a hashing algorithm. Another Blockchain solution, Docschain [3], was introduced to treat bockcerts' limitation platform under the consortium Hyperledger fabric. In addition to employing the hash function, Docschain utilized optical use character recognition (OCR) to extract the original data from the degree hard copies. Furthermore, Docschain integrated Internet of Things (IoT) devices using an IoT camera for executing the read operation. The CVSS [4] is another certificate verification platform implemented under the Ethereum Blockchain in Vietnam, where its functionalities are issuing, verifying, and retrieving student degrees.

In addition, researchers have proposed several blockchain-based degree issuing and verification approaches under the implementation of various blockchain platforms such as bitcoin [6,7], Ethereum [8,9], Hyperledger Fabric [10,11], multichain [12], [13], and tangle [14]. For more information about other works on the paper topic, the reader can refer to the systematic review [15] on integrating blockchain as a trust-distributed database for academic certificate verification.

In Algeria, great efforts have been made by the startup Takawen, which has attempted to fight forged training certificates using the Algerian certification and verification portal [5]. However, this experience does not apply in university degree certification, and the portal has solely settled for employing a traditional database system (Create/Read/Update/Delete), which is prone to attacks such as SQL-injection attacks, weak authentication attacks, and potentially privilege abuse. In this situation, blockchain presents an ideal solution to overcome these issues, although deploying blockchain across the entire national territory may face new challenges.

## 3. VerChain Architecture and its main Functionalities

Figure 1 depicts the VerChain components and their associated actors, showcasing all necessary smart contracts. The blockchain smart contracts proposed by VerChain enable users to engage with the BC network. The subsequent section offers comprehensive explanations of VerChain components and their authorized users, as well as their interactions.

### 3.1. Users Roles and Responsibilities

This section outlines the various import roles within VerChain. By breaking down these roles, the system enhances its security measures and clearly defines the responsibilities of each participant. This separate role contributes to the overall integrity and accountability of the system.

**Ministry of the Interior, Local Authorities and Regional Planning (Algeria):** Its main role is to manage blockchain data on persons who are under degree preparation. Each individual must be enrolled using the national identification number, which serves as a unique identifier. This distinct number can be utilized to connect every person to their blockchain-stored certificate. Due to the sensitive nature of citizen data, MILARP is the sole entity authorized for this task. MILARP has the capability to oversee access controls and identity management based exclusively on personal information.

**Minister of Higher Education and Scientific Research (Algeria):** Its main role is to register new certificate issuers, such as universities and institutes, because it is the main authority with this right. Every certificate issuer is registered using a unique ID along with its specific information. MHESR can manage access controls and identity management based only on certificate issuer data.

**Minister of Vocational Training and Education MVTE (Algeria):** It performs the same function as the MHESR, but lacks the authority to register certificate issuers in the fields of higher education and scientific research.

**Certificate Issuer:** This is the only authority that delivers the certificate. Its role is attributed to the Ministry of Education and Higher Education, or MVTE, by providing a blockchain certificate. It uses this certificate and the Algeria ID of a person to store the certificate degree in the blockchain.

**Certificate verifier:** This can be any organization that can verify the certificate degree of persons. For example, a company could verify the certificate degree of every candidate registered for a job.

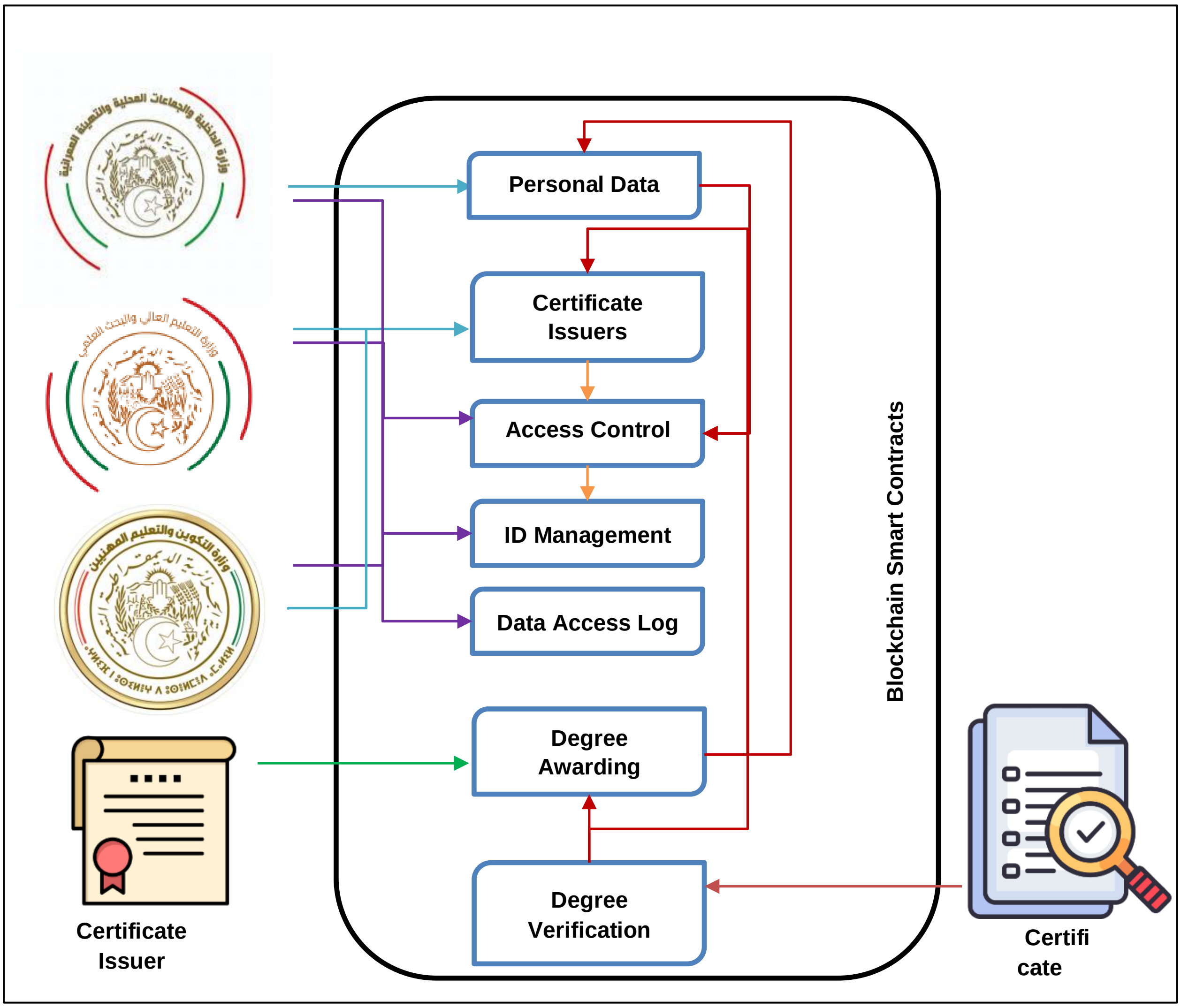


**Figure 1:** The Main Components of VerChain Architecture along with theirs Interactions

### 3.2. Blockchain Components

In this section, we provide descriptions of VerChain components where everyone is associated 0with one or several roles. Every component is considered to be a blockchain smart contract that ensures the services offered by that component. All components must interact with Access Control to verify whether the user has access authorization to use the target components.

**Personal data:** This manages data related to persons who have received certificate degrees from certificate issuers. These data are considered sensitive; therefore, we proposed that they be controlled and managed by the Minister of the Interior. Every person is registered with a unique Algeria ID.

**Certificate Issuers:** It uses MHESR or MVTE to control and manage the data related to certificate issuers. The MHESR or MVTE stores every registered issuer using its unique ID along with its information. Therefore, every unregistered issuer is not allowed to deliver a certificate. The MHESR or MVTE stores authorized degrees that can be delivered to every issuer.

**Identity Management**: It has two main tasks: creating identities and registering new VerChain users such as certificate issuers and verifiers, and verifying whether a given identity is valid using the blockchain. The IDM component returns a registration certificate for every accepted registration demand, which contains critical information about enrolling VerChain users with different roles. All identity information is stored in the BC to protect VerChain from fraudulent identities. Only MILARP, MVTE, and MHESR interacted with IDM.

**Access Control:** Access controls are defined for VerChain components, which are only used by MILARP, MHESR, and MVTE. For example, the BC administrator places restrictions on accessing the verifier component of the certificate. The restriction can be, for example, to provide access for a period of time to some verifiers.

**Access Log:** This stores all the operations performed by VerChain users. The main objective is to perform an advanced verification that detects inconsistencies in the history of authorization and access control components. It merges the data from all ledgers and performs advanced checks. For example, if a certificate issuer changes the data about a delivered certificate degree, this operation can be checked using audit verification. The auditing process is initiated by MILARP, MHESR, or MVTE.

**Degree Awarding:** This is invoked by a legitimate certificate issuer to register a new degree. In this situation, the certificate issuer and person ID must already be stored with their information using the certificate issuer data and personal data components, respectively. This restriction ensures that legal issuers deliver certificates to legal persons. Every certificate must be stored with a unique ID that can be, for example, a hash calculated using the certificate data.

**Degree Verification:** A legitimate certificate verifier uses this component to check whether a given certificate degree has been delivered to a person. The verification process uses information from the degree award to complete the check using the data provided by the certificate, such as the certificate hash and person ID.

## 4. Example of Scenario in VerChain and their Smart Contracts

In this section, we present an example of a scenario that can be controlled using VerChain smart contracts. Publishing and verifying a given certificate on the blockchain. ElTarf University takes the role of a certificate issuer, whereas Oracle Company plays a certificate verifier role. Mohamed an Algerian Citizen takes the role of a candidate requesting the Oracle Job. The following steps explain how this scenario can be executed.

1. Oracle requires candidates for a given job of a Data Scientist. Therefore, it requires a master's degree in computer science.

2. The MILARP registers Mohamed on the blockchain using his national ID along with his required personal data such as first name, last name, birthdate, address, etc.

3. After receiving a certificate of registration from the MHESR, ElTarf University issued a Master of Computer Science to Mohamed Using his national ID and his Personal Data. VerChain then checks whether Algerian University has the right to this kind of degree using onchain data on universities.

4. After registering Mohamed for the Oracle job using the hash of his certificate degree and national ID, Oracle uses its certificate automatically given by the VerChain administrator to verify the given degree through blockchain smart contracts.

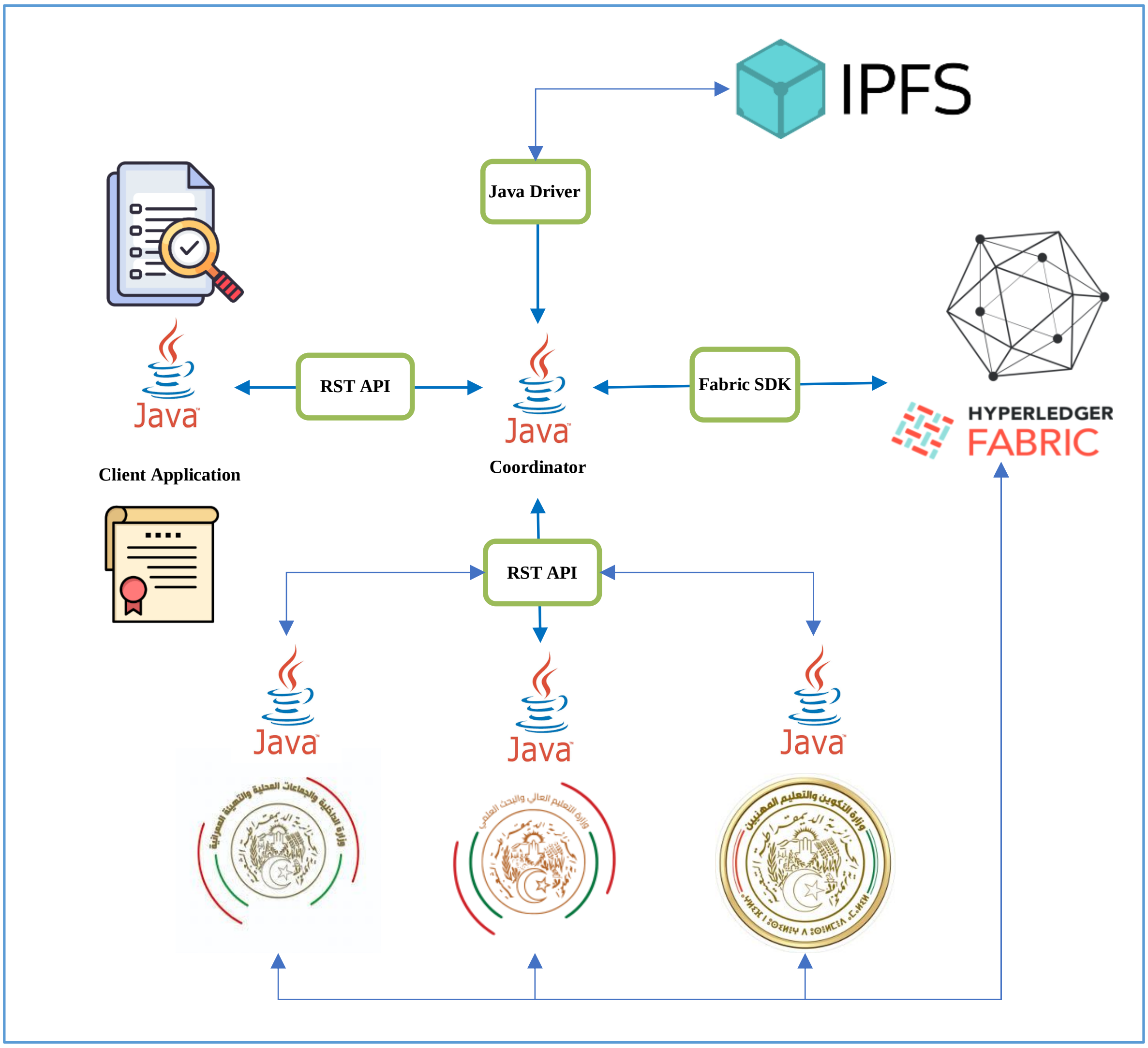


**Figure 2.** Implementation Architecture of the proposed VerChain.

## 5. Implementation

The proposed approach was implemented under Eclipse using various Java APIs, such as JSON and Fabric SDK. The implementation architecture is illustrated in Figure 2, where the main components are as follows:

*The Fabric Hyperledger Blockchain*[2] is used with the configuration of two organizations and one peer node for each one. The blockchain network uses CouchDB as a world state database and an ordering service. It uses a certificate authority for each organization. Six channels were created for identity management, degree warding, degree verification, Personal Data, Access Control, and logs. Six Fabric smart contracts were deployed using the Go language (one for each channel). The Hyperledger Fabric network used in VerChain is shown in Figure 3 where every channel is associated with its ledger and smart contract. Figure 4 illustrates the steps involved in executing a transaction in the fabric network for the chain code associated with the identity management channel.

[2] https://hyperledger-fabric.readthedocs.io/en/release-2.2/

*IPFS*[3]: The Interplanetary File System is a distributed file storage, where every file added to the IPFS is given a unique address derived from a hash of the file's content.

*Client Application*: a component that needs query results.

All these latter are interacted with the main program using their specific Java API.

### 5.1. Fabric Chaincodes and Distributed Ledgers in VerChain

Every peer in HLF has its local database (ledger), which contains all transactions executed by the network via HLF chaincodes. Thus, each peer can have several installed chain codes for a single HLF channel. The distributed ledgers in the HLF are updated by executing smart contracts with blockchain external users. Our work proposes the use of six distributed ledgers (one for every VerChain component), where each ledger is associated with one chaincode and several peers. These ledgers store critical data on VerChain operations, such as identity management degree awarding, degree verification, personal data, and operation logs.

#### 5.1.1. Identity Management

The IDM chaincode defines functions executed by HLF peers for managing the participants registered by the BC, such as certificate issues (e.g., universities), certificate verifiers (e.g., enterprises), and organizations Algerian Ministers. This chaincode is installed on a channel identified by the same name, Identity Management," and is associated with a local ledger that saves information about registered entities. The IDM chaincode uses the Golang structure, which is illustrated by *listing 1* whereas the IDM ledger uses the JSON key-value format given by *listing 2* as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the IDM ledger.

```
type Degree Issuer struct {
    DIsId string `json:"DISId"`
    DIsName string `json:"DIsName"`
    DIsType string `json:"DIsType"`
    DIsAddress string `json:"SIsAddress"`
}
type Organization struct {
    OrgId string `json:"OrgId"`
    OrgName string `json:"OrgName"`
    OrgType string `json:"OrgType"`
    OrgAddress string `json:"OrgAddress"`
}
```

**Listing 1.** The Golang Structures used by the Identity management Chaincode

[3] https://ipfs.io/

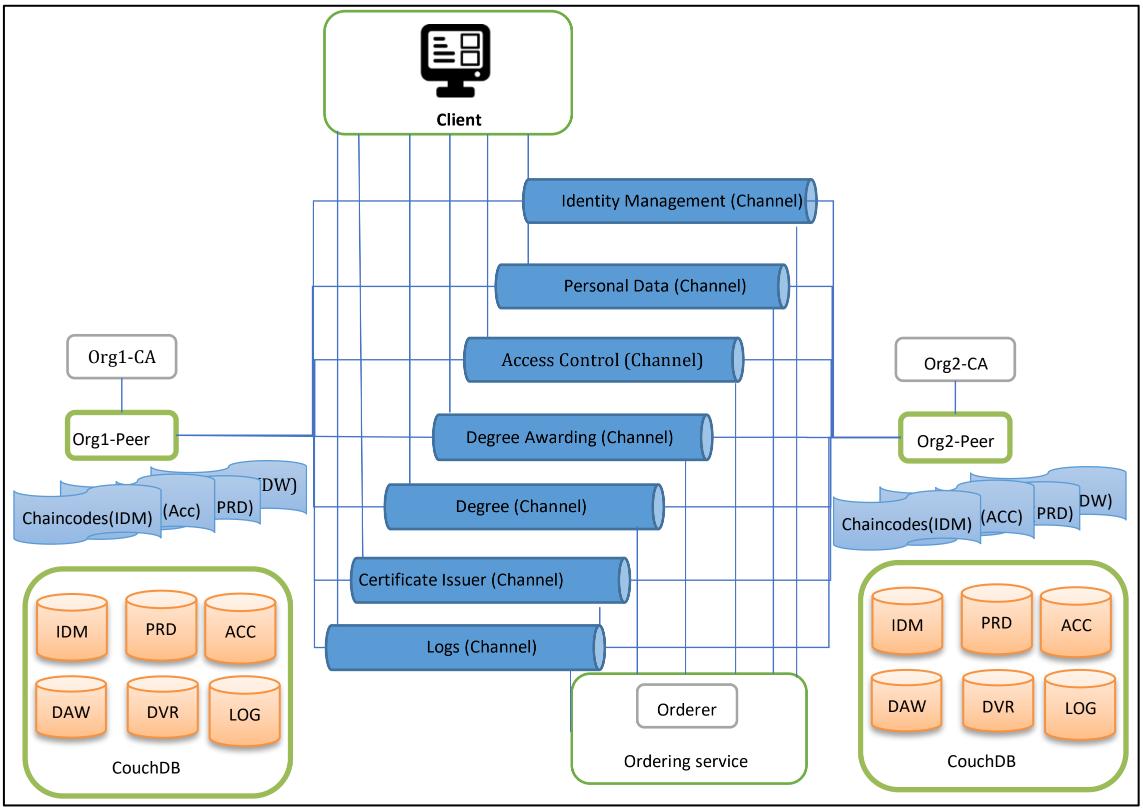


**Figure 3.** The Hyperledger Fabric Network Used by The Proposed Method

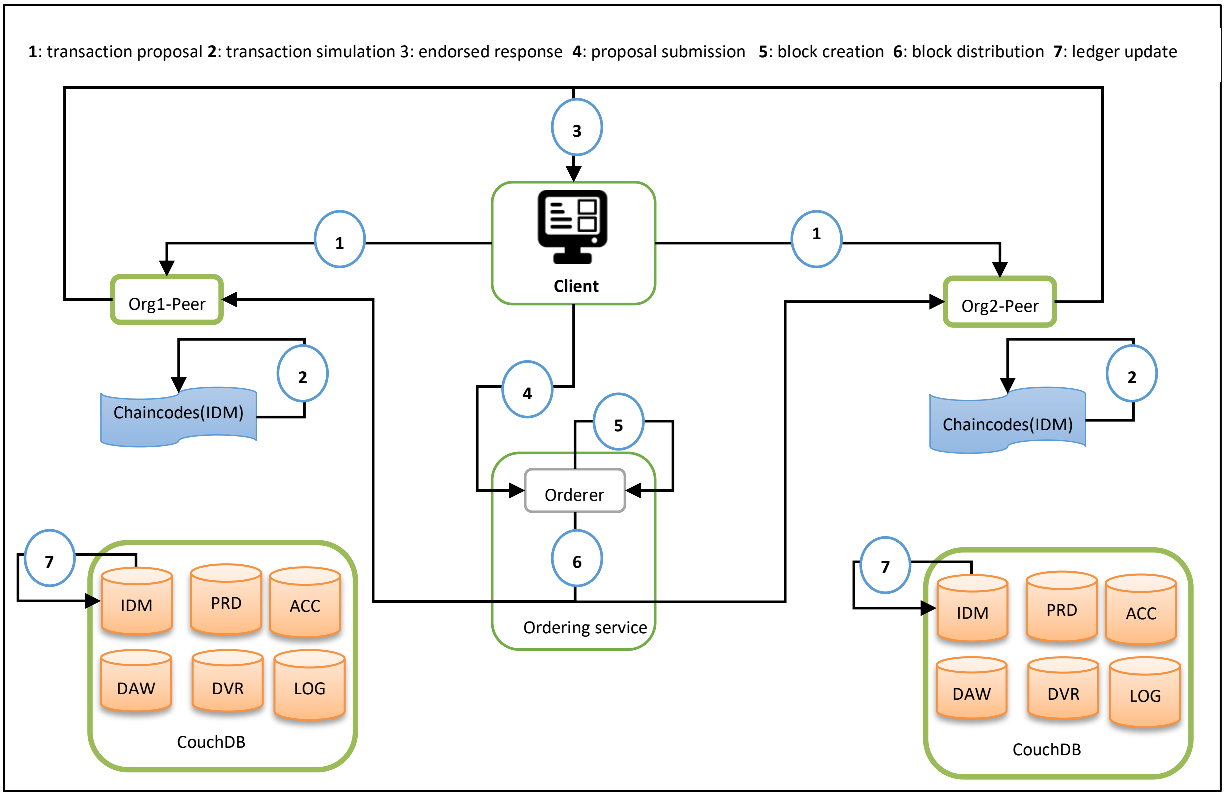

**1**: transaction proposal **2**: transaction simulation 3: endorsed response **4**: proposal submission **5**: block creation **6**: block distribution **7**: ledger update
3
1
Client
1
Org1-Peer
Org2-Peer
2
Chaincodes(IDM)
4
5
2
Chaincodes(IDM)
Orderer
6
Ordering service
7
IDM
PRD
ACC
DAW
DVR
LOG
CouchDB
7
IDM
PRD
ACC
DAW
DVR
LOG
CouchDB

**Figure 4.** Fabric Steps to Execute a New Transaction of the Chain-code Associated with Identity Management Channel

```
{"DIsID": "a21khjh44, ” ””,"DIsName":"National Training Center", "DIsAddress":"ElTarf",
"DIsType":"National Institution"}, {"OrgID": "a21kh32fr","OrgName":"ElTarf University",
"OrgAddress":"ElTarf", "OrgType":"University”}
```

**Listing 2.** Example of JSON key value Structure used by Identity management chaincode

| Function | Description | Restricted Access |
|---|---|---|
| CreateDegreeIssuer | Create new degree issuer using the description given by the invocation parameters | MHESR, MVTE |
| CreateDegreeVerifier | Register new degree verifier using the description given by the invocation parameters | MHESR, MVTE |
| GetAllCIs | Get all certificate issuers that are already registered by the BC | MHESR, MVTE |
| updateCI | Updating existed certificate issuer with new information. | MHESR, MVTE |

**Table 1.** Some Smart Contract Functions that are Implemented by identity management Chaincode

### 5.1.2. Access Control chaincode

The AC chain code defines the functions executed by HLF peers for managing access controls associated with registered participants. This chaincode is installed on a channel identified by the same name, Access Control,” and is associated with a local ledger that saves information about data access control and policies. The AC chaincode uses the Golang structure, which is illustrated by *listing 3*, whereas the AC ledger uses the JSON key-value format given by *listing 4* as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the AC ledger.

```
type Access Control struct {
        RequesterID string `json:" RequesterID "`
        Permissions [] PermissionType `json:" Permissions"`
        }
type PermissionType struct {
        Permission string `json:"Permission”`
        PermissionType string `json:" PermissionType"`}
```

**Listing 3.** An example of a Golang Structure used by the Access control Chaincode

```
{"RequesterID":"a21kh32fr", "Permissions":[{"Permission":"Issuing degree certificate",
"PermissionType":"Write":[{"Permission": "Verifying degree certificate",
"PermissionType":"Read"}]}
```

**Listing 4.** Example of JSON key value Structure used by the Access control Ledger

| Function | Description | Restricted Access |
|---|---|---|
| CreateACC | Create new access control for a given requester | MILARP, MHESR, MVTE |
| UpdateACC | Update a given access control for a given requester. | MILARP, MHESR, MVTE |
| GetACC | Get the access control associated with a given requester. | MILARP, MHESR, MVTE |
| RemoveACC | Remove an access control associated with a given requester. | MILARP, MHESR, MVTE |

**Table 2.** Some Smart Contract Functions that are Implemented by the Access control Chaincode

### 5.1.3. Personal data Chaincode

The PSA chaincode defines functions executed by BC peers for managing saved data about persons who have been awarded degrees. This chaincode is installed on a channel identified by the same name “Personal Data” and it is associated with a local ledger that saves information about awarding people which can be queried by members given by table 3. The PSA chaincode uses the Golang structure, which is illustrated by listing 5, whereas the SC ledger uses the JSON key-value format given by listing 6 as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the SC ledger.

```
type personal data struct {
        StudentNIN string `json:" StudentNIN"`
        StudentFirstName string `json:" StudentFirstName "`
        StudentLastName string `json:" StudentLastName "`
        StudentBirthDate string `json:" StudentBirthDate"`
        StudentEmail string `json:"StudentEmail "`
        StudentNumT string `json:"StudentNumT "`
        StudentAdr string `json:"StudentAdr"`
}
```

**Listing 5.** The Golang Structure used by the Personal Data Chaincode

```
{"StudentNIN ": "1099992515483625",
 "StudentLastName": "Mohammed",
 "StudentLastName": "Saadi",
  "StudentBirthDate": "12/9/2000",
 "StudentEmail": " Mohammed.Saadi@univ-elTaref.dz",
 "StudentNumT": "0655251649",
 " StudentAdr": "ElTarf, Algeria",
}
```

**Listing 6.** Example of JSON key value Structure used by the Personal Data Ledger

| Function | Description | Restricted Access |
|---|---|---|
| CreatePrData | Create new personal data on the blockchain. | MILARP |
| GetPrsData | Retrieve from the BC the personal data associated with a given NIN | MILARP |
| UpdatePrsData | Update an existing personal data for a given NIN | MILARP |

**Table 3.** Some Smart Contracts Functions that are Implemented by the Personal Data Chaincode

```
    type Degree Awarding struct {
        StudentNIN string `json:" StudentNIN"`
        DegreeID string `json:"DegreeId"`
        DegreeDeliverID string `json:"DegreeDeliverID"`
        DegreeTitle string `json:" DegreeTitle"`
        DegreeType string `json:" DegreeType"`
        DegreeSpeciality string `json:" DegreeSpecialty"`
        DegreeMention string `json:" DegreeMention"`
        SignedBy [] SignatureType `json:"SignedBy"`
    type SignatureType struct {
        PersonNIN string `json:" PersonNIN”`
        PersonQuality string `json:" PersonQuality"`}
    }
```

**Listing 7.** The Golang Structure used by the Service Composition Chaincode

```
{"StudentNIN ": "1099992515483625",
 "DegreeDeliverID": "a21kh32fr",
 "DegreeID": " as1kh2562",
  "DegreeTitle": "Certificate issuing and verification using Blockchain",
 "DegreeType": "Master of Engineering",
 "DegreeSpeciality": "Information Security",
 "DegreeMention": "Very Good",
 "SignatureType":[
   {“PersonNIN”: "1099260215022518”,
   “PersonQuality”: "Dean of the faculty of Sciences and Technologies”},
   {“PersonNIN”: "1099512502160217”,
   “PersonQuality”: "Rector of Sciences and Technologies”}
]}
```

**Listing 8.** Example of JSON key value Structure used by the Service Composition Ledger

### 5.1.4. Degree Awarding Chaincode

The DAW chaincode defines functions executed by BC peers for managing saved data about degree- awarding processes, such as degree delivery institutions and target students. This chaincode is installed on a channel identified by the same name "Degree Awarding" and it is associated with a local ledger that saves information about the degree awarding process which can be queried by members given by table 4. The SC chaincode uses the Golang structure, which is illustrated by listing 7, whereas the SC ledger uses the JSON key-value format given by listing 8 as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the SC ledger.

| Function | Description | Restricted Access |
|---|---|---|
| GetDegreeAW | Get a degree data using its DegreeID. | MHESR, MVTE and theirs registered degree awarding organizations. |
| CreateDegreeAW | Register a new degree awarding data. | Registered degree awarding organizations |
| UpdateDegreeAW | Update an existing degree data | Registered degree awarding organizations |

**Table 4.** Some Smart Contracts Functions that are Implemented by the Service Composition Chaincode

## 6. Evaluation

### 6.1. Experiment Configuration

To validate the functionality and test the performance of our approach, several experiments were performed. Experiments were performed on a machine with an Intel Core i7 processor running with a 1.8 GHz clock speed, 16 GB memory, 128 GB SSD, and 1 TB for storage. The components of the fabric network were deployed as Docker 2.3 images (organizations, certificate authorities, peers, CouchDB. etc.). With regard to the implementation architecture, the coordinator is implemented as a JAVA REST application that uses Tomcat 9 as the resource server. All service ministries and clients (certificate issuers and verifiers) are depicted as JAVA applications that communicate with servers using the REST API. Our implementation uses several Java API in different processes, such as the REST API, IPFS API, and Fabric SDK.

## 7. Conclusion

This paper proposes VerChain, an Algerian platform blockchain-based system for the generation and verification of academic degrees and student credentials. First, the proposal presents the existing solutions of several blockchain implementations that share similar endeavors. Subsequently, it sheds light on Algerian architecture that detailed the interaction between certificate verifiers and government ministries, along with their role and access restrictions. Moreover, the study underlined the privacy of data related to students' credentials, issuers, and even certificates using access control. Therefore, verchain-managed access is allowed only by a valid identity and a specific role. VerChain ensures automatic certificate issuing and verification, which limits errors and omits manual processes. Traceability is guaranteed through the log component to start the auditing process to detect any error or illegal access, and the communication between clients and VerChain uses TLS for better security and privacy preservation.

## References

[1] ANPDP: https://anpdp.dz/fr/accueil/.

[2] P. Schmidt. (Oct. 24, 2016). Blockcerts—An Open Infrastructure for Academic Credentials on the Blockchain MLLearning, MIT Media Lab. [Online] Availble: https://medium.com/mit-media-lab/blockcertsan- open-infrastructure-for-academic-credentials-on-the-blockchain-899a6b880b2f

[3] S. Rasool, A. Saleem, M. Iqbal, T. Dagiuklas, S. Mumtaz and Z. u. Qayyum, "DocsChain: blockchain-Based IoT Solution for Verification of Degree Documents", IEEE Transactions on Computational Social Systems, vol. 7, no. 3, pp. 827-837, June 2020.

[4] Nguyen, D. H., Nguyen-Duc, D. N., Huynh-Tuong, N., & Pham, H. A. (2018, December). CVSS: a blockchainized certificate verifying support system. In Proceedings of the 9th international symposium on information and communication technology (pp. 436-442).

[5] Algerian Certification & Verification Portal access: https://www.takawen.dz/about-us-2/.

[6] M. Sharples and J. Domingue, “The blockchain and kudos: A distributed system for educational record, reputation and reward,” in Proc. Eur. Conf. Technol. Enhanced Learn. Cham, Switzerland: Springer, 2016, pp. 490–496.

[7] A. S. de Pedro Crespo and L. I. C. García, “Stampery Blockchain Timestamping Architecture (BTA)–Version 6,” 2017, arXiv:1711.04709. [Online]. Available: http://arxiv.org/abs/1711.04709

[8] T. Kanan, A. T. Obaidat, and M. Al-Lahham, “SmartCert blockchain imperative for educational certificates,” in Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT), Apr. 2019, pp. 629–633.

[9] B. Boeser. (Jul. 2017). Meet Truerec by Sap: Trusted Digital Credentials Powered by Blockchain | Sap News Center. Accessed: Jan. 29, 2020. [Online]. Available: https://news.sap.com/2017/07/meet-truerec-by-saptrusted-digital-credentials-powered-by-blockchain/

[10] “A permissioned blockchain-based system for verification of academic records,” in Proc. 10th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS), Jun. 2019, pp. 1–5.

[11] J. A. Otuya, “A blockchain approach for detecting counterfeit academic certificates in Kenya,” M.S. thesis, Dept. Inf. Technol., Strathmore Univ., Nairobi, Kenya, 2019.

[12] G.-A. Dima, A.-G. Jitariu, C. Pisa, and G. Bianchi, “Scholarium: Supporting identity claims through a permissioned blockchain,” in Proc. IEEE 4th Int. Forum Res. Technol. Soc. Ind. (RTSI), Sep. 2018, pp. 1–6.

[13] R. Arenas and P. Fernandez, “CredenceLedger: A permissioned blockchain for verifiable academic credentials,” in Proc. IEEE Int. Conf. Eng., Technol. Innov. (ICE/ITMC), Jun. 2018, pp. 1–6.

[14] A. Wahab, M. Barlas, and W. Mahmood, “Zenith certifier: A framework to authenticate academic verifications using tangle,” J. Softw. Syst. Develop., vol. 2018, no. 370695, p. 13, 2018.

[15] A. Rustemi, F. Dalipi, V. Atanasovski and A. Risteski, "A Systematic Literature Review on Blockchain-Based Systems for Academic Certificate Verification," in *IEEE Access*, vol. 11, pp. 64679-64696, 2023, doi: 10.1109/ACCESS.2023.3289598.

# Unmanned Aerial Vehicles Security: Threat Analysis and Hybrid Cryptography for secured Communication

1st Zina Oudina
*Embedded Systems Laboratory*
*Badji Mokhtar Annaba University*
Annaba, Algeria
zina.oudina@univ-annaba.org

2nd Chaouki Chemam
*LIMA Laboratory*
*Badji Mokhtar Annaba University*
Annaba, Algeria
chemam-chaouki@univ-eltarf.dz

***Abstract*—Modern Unmanned aerial vehicles (UAVs) can be utilized for various civilian applications, including traffic surveillance, weather updates, agriculture, photography, firefighting, and product delivery. The military has recently increased the use of UAVs for essential operations to decrease the exposure of their important human resources in areas of greatest risk. UAVs consist of three major modules: the data transmission link, the aircraft, and the ground control base station. The UAV and GCS typically communicate via several protocols, such as UranusLink, UAVCan, and Micro Aerial Vehicle Link (Mavlink). UAVs are subject to a variety of attacks that might disrupt communication, including spoofing, denial of service and packet attacks, GPS attacks, and others. The security threats are various, and as attacker techniques are increasingly developed to target UAV communication protocols, existing vulnerabilities may be exploited by the attackers. This paper investigates and analyses vulnerabilities in UAV communication protocols, and proposes hybrid cryptography methods to mitigate those threats.**

***Index Terms*—Unmanned aerial vehicles (UAVs), Threats, Risks, Communication, Micro Aerial Vehicle Link (Mavlink), Hybrid Cryptography.**

## I. Introduction

Unmanned Aerial Vehicles (UAVs) can fly without a human pilot and can be operated remotely using radio or WiFi. Numerous civilian uses for modern UAVs include product delivery, agriculture, photography, firefighting, traffic surveillance, and weather updates [1]. Three primary parts comprise UAVs: the aircraft, the ground control base station, and the data transmission link.

A group of one or more computers with a video screen attached is called a ground control station (GCS) and is used to oversee UAV flight operations [2]. The unmanned aerial vehicle (UAV) is tracked by the GCS using information gathered from several sensors mounted on the vehicle to determine its position, velocity, altitude, and mission status. The UAV and the ground control station are linked via a communication link. WiFi, Bluetooth, 4G, 5G, and 3G are among the UAV communication types [3].

Typically, the Micro Aerial Vehicle Link (Mavlink) [4], which is the most used protocol, UranusLink [5], UAVCan [6], and other protocols are used for communication between the UAV and the GCS. A graphical user interface (GUI) or command-line interface, a wireless datalink subsystem, a processor unit, a telemetry/telecommand module, a user control module, and a wireless datalink subsystem make up a GCS.

The configuration of a wireless datalink subsystem enables remote communication with a UAV [7]. The telemetry/telecommand module is set up to send instructions from the ground station to the UAV and to download data from the UAV while it is in flight. The display module that makes up the user interface is set up to show downloaded data from UAVs.

### A. Literature review

The research [8] addressed risk mitigation, categorized UAV security concerns as internal and external threats, and offered practical advice on how to create a preventive plan to guard against drone vulnerabilities.

In [9], they investigated the viability of implementing hijacking and jammer attack vectors as means of interfering with a current flying operation. The outcomes of the experiment show how successful these attacks may be when using a Crazyflie 2.1 drone in both autonomous and non-autonomous flight modes. Additionally, they offered possible shielding tactics that would ensure a safe and secure flight. Different cryptographic protocols have been applied to protect communication between UAVs. The cryptography methods required to establish a successful UAV communication have been covered in the articles [10], [11].

A multi-tier adaptive military MANET security protocol employing signcryption and hybrid cryptography is presented in the study [12]. The protocol helps with innovations to secure military MANET connections for three primary reasons: the military MANETs' structural organization, hybrid key management protocols, and cryptographic techniques. When compared to several conventional cryptographic techniques, they employed the Elliptic Curve Pintsov-Vanstone Signature Scheme (ECPVSS), which offers security and performance benefits.

A. Allouch et al. [13] described the security and vulnerability risks to the MAVLink communication protocol and recommended using a number of cryptographic algorithms (RC4, AES Counter mode, AES Cipher Block Chaining Mode, and ChaCha20) to mitigate these risks. They also assessed how well these algorithms performed in terms of CPU usage and memory utilization. In comparison to other cryptographic algorithms, their experiments showed that the lightweight ChaCha20 method was highly effective.

### B. *Contribution*

When the communication and encryption protocols employed by UAVs are weak and inefficient, the UAV becomes increasingly vulnerable to a variety of assaults. Attackers can easily target UAVs by interfering with their connectivity with remote locations and centers. Wireless connections between UAVs make them vulnerable to security threats such as data tampering, denial of service, spying, dispatch systems, ADS-B, man-in-the-middle, and WiFi attacks. Unmanned Aerial Vehicles (UAVs) represent a substantial risk of damage; hence, it is critical to understand the security and safety implications. Attackers can simply take control of a UAV using standard hacking tools, stopping it from carrying out its intended functions or, worse, causing substantial damage.

This article analyzes vulnerabilities in UAV communication protocols. The weaknesses and attacks are detailed, as well as the tools and scenarios employed, which vary in approach and strategy.These assaults include man-in-the-middle attacks, denial-of-service (DoS), sniffing attacks, GPS spoofing, and eavesdropping attacks.As a solution to communication attacks, this study presents a hybrid cryptography for safeguarding the UAV communication protocol.

The aims of this paper are:

- Sheds some light on current security challenges and vulnerabilities in UAV communication.
- Analyzing attack techniques and scenarios.
- Explaining how cryptography protects data during UAV communications and suggesting a hybrid encryption technique.

### C. *Paper structure*

This study is divided into five sections, with Section 2 addressing security threats and vulnerabilities in UAV communication. The most commonly used protocols for UAV communication are explained, as is an analysis of attack techniques and scenarios. Section 3 describes cryptography for UAV communication and proposes a hybrid cryptography for secure UAV communication. Section 4 contains the discussion, and the paper concludes in Section 5.

## II. SECURITY THREATS AND VULNERABILITIES OF UAV COMMUNICATION

This section focuses on the weaknesses and vulnerabilities in UAV communication, as well as the common attacks that threaten their security. Integrity, availability, secrecy, privacy, and authenticity are key security requirements for UAV communication.

- Integrity: integrity denotes the preservation and unaltered state of the original conveyed signals and information.
- Availability: The system can function at the user's convenience when required and accessibility to the service ought to be offered whenever the user needs it.
- Confidentiality: Only authorized users can access the system's information.

### A. *Vulnerabilities in UAV communication*

The main vulnerabilities for UAV communication come from the lack of security design in UAV systems and UAV networks and the used communication protocol. An attacker can interfere with UAVs in different ways, including taking control and disabling communication between UAVs and the ground control station (GCS) as shown in Figure 1. The lack of information cryptography in UAV protocol encourages many violations that target UAV communication such as: a) Integrity violation: replay attacks, the injection of the message. b) Availability violation: can be done with denial of service attacks. c) Confidentiality violation: can result in information eavesdropping and system spoofing.

Many vulnerabilities threaten UAV systems, like malicious UAV detection, vulnerabilities in packet forwarding and routing protocols, jamming attacks, WiFi insecurity, false data injection, GPS spoofing, attacks on routing Protocols, and others. The privacy of the UAV network can be maintained by safeguarding the identities of each UAV in the network as well as the identity of the GCS. An attacker can compromise and use the identity of one UAV to connect with the network.

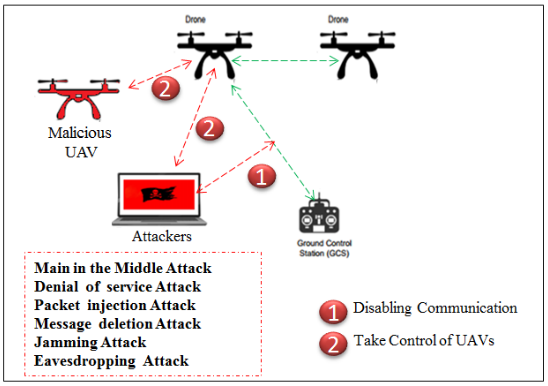


Fig. 1. Vulnerability for UAV communication.

### *B. Security threats and commonly used protocols in UAVs Communication*

UAVs face not just security concerns but also insufficient resources such as battery life, real-time computing, and autonomous control. Existing protocols are vulnerable to numerous dangers due to limited energy resources, communication bandwidth, and processing capacity.

The most popularly used protocol is MavLink. (Micro Air Vehicle Link). MAVLink [14] protocol is a lightweight open source protocol used to remotely operate unmanned aerial vehicles (UAVs) by messages transmitted from a ground-operated station (GCS) and for bidirectional communications between cooperative UAVs. This protocol provides robust features for tracking and managing UAV missions.

Most autopilot systems include MAVLink, primarily ArduPilot and PX4. MAVLink is available in two versions: 1.0 and 2.0. The MAVLink 2.0 protocol, which is the recommended version and was introduced in 2017, replaced the 2009-issued MAVLink 1.0 protocol.

The security level is higher with MAVLink 2.0 than it is with MAVLink 1.0. Three components make up an MAVLink 2.0 packet: a header with information about the message; a payload with data carried out by the message; and a trailer with a checksum to ensure that the message is not altered during transmission and a signature to confirm the integrity of the data and the legitimacy of the message's source (originating from a trust node). The length of an MAVLink packet might vary, ranging from 11 bytes to 297 bytes, contingent upon the characteristics that are shared.

### *C. Analysis of attack techniques and scenarios*

An attack scenario varies in its approach and strategies. One common scenario involves intercepting UAV network communications and injecting packets to disable the UAV. The UAV system is connected via a network, and the attacker gains access to the network using various current methods. The following are the most commonly employed attacks:

- Man-in-the-middle attack: The man-in-the-middle attack is an act permitting unauthorized access to private information and intercepting and altering communications between two UAV parties in a network. MITM attacks jeopardize the security and integrity of data flow and come in a variety of ways, including packet sniffing, spoofing the DNS, and session hijacking.

  An attacker positions himself between the drone and the GCS to sniff information. An attacker attempts to connect to the drone's controller. After gaining access to the drone, the opponent sends counterfeit commands to the controller, deceiving them into believing they are communicating with the drone. That allows all data sent from drones to GCS and vice versa to be intercepted by the attacker. A man-in-the-middle attack compromises also the confidentiality of the UAV systems [15].
- Denial-of-Service (DoS) attacks: DOS attacks prohibit users from accessing system services [16], target available resources, and employ logic and resource attacks [17]. Through a deluge of requests, the attacker can overload the flying UAV's resources and reduce its availability by flooding the network card with random traffic. [18] Another way to initiate a denial-of-service attack is to try to deactivate the control signal by sending large numbers of packets to the GCS within a predefined range. This results in the drone experiencing a lost connection, which ultimately leads to a malfunctioning data link.

  The most frequent and simple kind of attack is a denial of service (DoS), in which the attacker overloads the system by sending a large number of requests or packets. When the attacker delivers the drone multiple data packets, the drone's computing power fails and the network link between the drone and the ground controller is de-authenticated [19].Another possibility is that one of the transmitted data packets contained malicious code that might be used to take down the drone. The hijacker may utilize such attacks to bring down the drone and hurt citizens and government organizations.

  One kind of denial-of-service attack is the de-authentication attack, which interferes with client-to-Wi-Fi access point communication. In this attack, the attacker de-authenticates the ground pilot, therefore denying the pilot control of the drone. Because he simply needs the drone's Mac address, which is made public by any program like ”Aircrack-ng” [20], which can hijack the drone, the attacker can transmit a de-authentication frame to a wireless access point at any moment. The connection between the drone and the ground controller is de-authenticated as soon as the Aircrack-ng utility is turned on. This tool allows the attacker to communicate with the drone and give it hostile instructions.
- Sniffing attacks : In a sniffing attack, an attacker monitors the communication channel without the participant's knowledge or consent and has the ability to gather sensitive data. This can be crucial even when using an encrypted communication link, since an attacker may try to intercept the communications by exchanging cryptographic keys to obtain the key and decrypt the data.
- Traffic Injection attacks: In a traffic injection attack, the attacker inserts their communication into the channel. Such communication could take the form of spoof messages sent to the relevant systems or a more subtle change in the existing communication. Traffic injectiTraffic Injection attacks: In a traffic injection attack, theon attacks aim to impair channel quality, interrupt information flow (availability), and spread misleading information.
- Spoofing attacks: The spoofing attack is employed to

intercept communications, possibly alter them, and then send them to the intended recipient without alerting them to the communication's interception.

- Identity spoofing attack: The majority of UAVs lack encryption safeguards; therefore, an attacker can employ identity spoofing to enter a communication channel and assume the role of a third party in order to gain the system ID of a UAV [21].
- GPS Spoofing: Using several transmitting antennas, a spoofing attack can be carried out. In this scenario, the attacker's broadcasting antenna combines with the matching receiving antenna to transmit false signals. Drones with no encryption on their chipboard are easily traced by hackers, and they share an incorrect location with the drone controller using a directional antenna with a narrow beam width aiming for the drone. Military drones have sophisticated encryption systems installed, making it challenging to spoof them. For generating fake GPS signals, it used a software-defined radio (SDR) and a BladeRFx40 SDR.
- Eavesdropping attacks: A transmission could be intercepted by an unauthorized party. The attacker then breaches the network's privacy by listening in on the discussion without stopping the transmission and has the power to add or remove messages [22].

Table 1 presents the description, techniques, and impact of some communication attacks.

## III. CRYPTOGRAPHY FOR UAV COMMUNICATION

### *A. Cryptography for UAV communication*

Encryption is the most often used technology for securing transmission and communication for UAV systems. Many types and methods of cryptography that address the security of UAVs exist in literature and hybrid cryptography is a novel line.

A novel hybrid cryptographic framework for secure data storage in cloud computing: Integrating AES-OTP and RSA with adaptive key management and Time-limited access control is proposed in [23].

The study introduced an intelligent framework for key creation, distribution, and rotation, progressively enhancing the security of cryptographic operations. In the study [24], it is suggested that a hybrid cryptography algorithm be used to send and receive data securely. The technique is examined in terms of throughput, encryption and decryption times, and security analysis using the avalanche effect.

The authors of [25] offered drone positioning optimization using IMU and UAV communication security with hybrid cryptosystems. A robot operating system (ROS) simulation environment is used to introduce a strategy of employing inertial measurement unit (IMU) data for location optimization. A hybrid cryptosystem that includes signature authentication using the RSA technique built- into MATLAB and Python3 protects the sensitive IMU values of UAVs used for communication.

TABLE I
DESCRIPTION AND TECHNIQUES; IMPACT OF SOME COMMUNICATION ATTACKS

| Attack | Description | Objectives | Impact |
|---|---|---|---|
| Main in the Middle | -Intercept link<br>- Uses a device Wi-Fi Pineapple<br>–Sends fake commands | Availability | Alter the network link |
| Denial of Service | Using superfluous requests-Restricted shared resource to legitimate users<br>-Sends several data packets to the drone , | Availability | -Failure the computational power of drone<br>- Make the network of drone overflow<br>- loss drone control |
| Eavesdropping attacks | -Listen discussion<br>- Add or remove message | Privacy | - Intercept the communication<br>- Permit unauthorized party. |
| Traffic Injection | Inject false information, fake commands | -Reduce the quality channel.<br>- Disrupt the flow of information (availability)<br>-Inject fake message | - Fake message with wrong commands to influence negatively UAV system decision. |
| Spoofing | Intercept, modify message | Availability, Integrty | Fake communication |
| Spoofing ID system | Employ identity spoofing , Enter communication channel. | Impersonating the UAV identity | Fake communication |
| GPS Spoofing | - Using transmit antennas<br>-Transmits false signal | Availability | -Reduce the velocity drone, making it less useful. |

### *B. Proposed hybrid cryptography for secured UAV communication*

MAVLink is one of the most commonly used protocols for UAV because of its differentiating feature; nonetheless, it does not perform any encryption and provides no security to the payload or messages. To keep hackers away from personal data and to ensure that a permitted drone is being controlled by GCS, it is crucial to verify the legitimacy of the unmanned systems via GCS.

This paper proposes using hybrid encryption to encrypt the identity IDs of GCS and the identity IDu of current UAVs in the same network. An initial IDs is provided for GCS and IDui for UAV by the operator. A key generated also for each identity. The hybrid encryption for UAV ID and GCS ID proceeds via two encryption phases, the first is realized by SHA256, and the second by SHA3-224 as shown in Figure 2. The basis of our proposal is to enable only encryption without decryption.

The first encryption is provided in preparation for the start of the UAV network mission. Each UAV will get the initial ID, which will be the identifier of each system as well as the GCS

TABLE II
THE PROPOSED HYBRID CRYPTOGRAPHY

| Process of proposed Hybrid Encryption |
|---|
| i=n, number of UAV |
| IDu: Initial ID provided from the operator to UAV |
| IDs: Initial ID provided from the operator to GCS |
| 1- Generate key for each ID |
| 2- Encrypt initial ID for UAV and GCS |
| 3- Encrypt the ID provided from Encryption 1 |
| 4- Storage of Encrypted ID for UAVs and GCS for all systems in UAV network. |
| 5- Send message form the GCS or the UAV Sender with initial ID |
| 6- Receive message from GCS of Receiver UAV |
| 7- Encrypt the received initial ID using (key of encryption1) |
| 8- Encrypt the ID result of encryption1 in (encryption key of encryption2) |
| 9- Verify the encryption of received ID with stored ID if the same, the UAV receiver will accept message and interact with sender |

and will be sent in a Mavlink2 packet. Those initial IDs that are provided by the operator will be encrypted with the proposed hybrid encryption, and the resulting encrypted ID will be stored in all databases of UAVs and GCS in the network. When the mission starts, the Mavlinks2 packet is sent to the UAV receiver. The latter will receive the packet and encrypt the received ID (encryption 1). The result of encryption 1 will be encrypted (encryption 2). The final encrypted ID will be compared with the stored ID, and the UAV receiver will confirm if the communicator is legitimate and is a member of the UAV network. The proposed hybrid cryptography is presented in Figure 2, and the steps of the process are in Table 2.

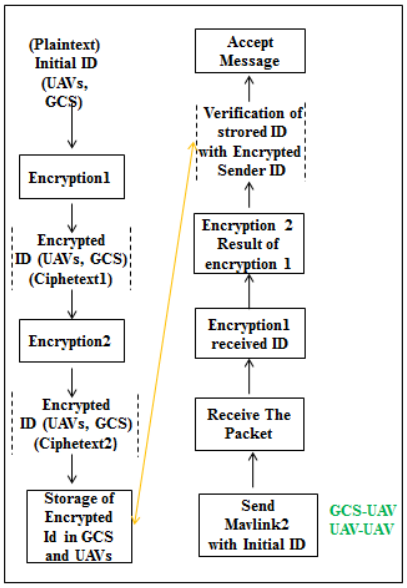


Fig. 2. The process of proposed hybrid cryptography

### *C. Case Study*

In this section, we look to prove the feasibility of our proposed hybrid cryptography and how it mitigates existing attacks for UAV communication.

TABLE III
THE RESULT OF THE SECOND STAGE, WHICH IMPLEMENTS ONLY ONE ENCRYPTION

| Initial ID | Encrypted ID Size Bytes | Encryption Time S |
|---|---|---|
| 'ID82' | 103 | 0.476 |
| 'ID77' | 104 | 0.352 |
| 'ID93' | 104 | 0.343 |
| 'IDGCS' | 104 | 0.358 |

TABLE IV
THE RESULTS OF STAGE THREE, WHICH IMPLEMENTS HYBRID CRYPTOGRAPHY

| Initial ID | Encrypted ID Size Bytes | Encryption Time S |
|---|---|---|
| 'ID82' | 189 | 0.682 |
| 'ID77' | 187 | 0.662 |
| 'ID93' | 187 | 0.610 |
| 'IDGCS' | 188 | 0.642 |

We try to implement the proposed hybrid cryptography by providing an initial ID to GCS (IDGCS) and an initial ID to UAVs (ID82, ID77, and ID93). For the first stage, we provided only one key for all the IDs, and we implemented only one encryption algorithm, SHA3-224. We noticed that the execution time of encryption is very small, but the encrypted ID is low in terms of privacy because the generated key is the same. In the second stage, we generated for each initial ID one key and implemented encryption 1 (algorithm SHA3-224) only, and the results are presented in Table 3. The third stage implements hybrid encryption for the initial provided ID of UAVs, and the results are presented in Table 3.

The size of the encrypted ID is considered in stage three, but we noticed that the size is not different or big due to the means of identification, which is not a message that contains a lot of details. The result of stage three is presented in Table 4.

In our proposed hybrid cryptography, we do not consider decryption. When it saves the encrypted ID and key in the early stages and sends the initial ID that protects the information within the transmission and even an attack is done, such as main in the middle, eavesdropping, or ID spoofing and sniffing, the attacker can't know the stored one, as shown in Table 5.

TABLE V
THE ADVANTAGES OF THE PROPOSED HYBRID ENCRYPTION

| Attacks | Proposed Hybrid encryption Protection |
|---|---|
| Main in the middle | Even the transmission is altered; the sent ID is the initial one and is not used for verification and authentication. |
| Eavesdropping | The encrypted ID and related encryption key are already saved in the receiver and no need to send the encryption and decryption keys , eavesdropping can't get keys. |
| Id spoofing | The sent ID in transmission is the initial one, even is spoofed it will be not used for authentication. |
| Sniffing | An attacker can't obtain the key and decrypt the data. |

## IV. DISCUSSION

The UAV systems are a means of cyber security tools if they are used as a testbed [26], and a means of threats and harm if they are used for negative purposes [27]. The most used attack and the easiest one is DOS attacks. Restrictions apply to authorized users' access to shared resources. Due to the overburden this may create on the system, some or all valid requests may not be fulfilled.

The attacker sends the drone many data packets during this process, de-authenticating the drone's network link with the ground controller and causing the drone's processing power to fail. The drone's network will overflow due to the rapid packet transmission rate, which will cause the drone to lose control over both itself and the ground controller. For the other attacks, we can reveal some remarks:

- There is a very clear synergy between communication attacks and software attacks for UAV systems.
- An attacker may merge between many types of attacks to break the UAV communication; sometimes the man-in-the-middle attack is the start point of a series of attacks followed by eavesdropping that violates the confidentiality of UAV systems.
- The use of spoofing attacks and traffic injection allows for the total takeover of unmanned devices with some degree of autonomy.

Knowledge of attack types and awareness is a pillar of defense strategy [28]. Encryption is the most often used technology for securing transmission and communication in UAV systems, and more security methods are suggested in many industries [29].

The hybrid encryption technique is proposed in this paper and merges two hash algorithms to ensure confidentiality, integrity, and authentication. Data encryption is made using a dynamically generated key. Two algorithms, SHA256 and SHA3-224, are used to encrypt the system ID in the UAV network, and the generation of keys is also used. The hash is used especially for the assurance of integrity and the propriety of one-way encryption without decryption to increase the security of an encrypted ID.

Some steps are required to complete the suggested hybrid encryption, such as simulation and some realistic scenarios. An authentication system in the UAV network system is also required. The authentication system will provide additional security lines for communication.

## V. CONCLUSION

The vulnerabilities in UAV communication protocols are examined and analyzed in this paper, along with the techniques, scenarios, and effects of various attacks, including man-in-the-middle attacks, Denial-of-Service (DoS) attacks, sniffing attacks, traffic injection attacks, spoofing attacks, GPS spoofing, and eavesdropping attacks.

Cryptography is a unique way of protecting transmission messages between UAVs -GCS and UAV-UAV. This paper proposes hybrid cryptography for protecting the identity of UAV and GCS and avoiding many attacks. This technique combines two different hash algorithms to assure confidentiality, integrity, and authentication. Data encryption is performed with a dynamically generated key. The UAV network uses two algorithms to encrypt the system ID: SHA256 and SHA3-224. The hash is used to ensure the integrity and propriety of one-way encryption without decryption, hence increasing the security of an encrypted ID.

Many issues exist for the security of UAV communication protocols, even when encryption is offered. The increasing range of attacks and technologies necessitates the development of many security levels using various encryption algorithms and a secure authentication mechanism. Our next work will be focused on implementing a multi-level security and authentication system for the UAV communication protocol.

## REFERENCES

[1] Hiebert B, Nouvet E, Jeyabalan V, Donelle L. The application of drones in healthcare and health-related services in north america: A scoping review. Drones. 2020 Sep;4(3):30

[2] Alladi T, Bansal G, Chamola V, Guizani M. Secauthuav: A novel authentication scheme for uav-ground station and uav-uav communication. IEEE Transactions on Vehicular Technology. 2020 Oct 22;69(12):15068-77.

[3] Islam N, Rashid MM, Pasandideh F, Ray B, Moore S, Kadel R. A review of applications and communication technologies for internet of things (Iot) and unmanned aerial vehicle (uav) based sustainable smart farming. Sustainability. 2021 Jan;13(4):1821.

[4] N. A. Khan, N. Jhanjhi, S. N. Brohi and A. Nayyar, "Emerging use of UAV's: Secure communication protocol issues and challenges," in Drones in Smart-cities: Security and Performance,1st ed., vol. 1. pp. 37–55, Chap. 3, Sec. 1. Turkey: Elsevier, 2020.

[5] V. Kriz and P. Gabrlik, "Uranuslink-communication protocol for uav with small overhead and encryption ability," IFAC-PapersOnLine, vol. 48, no. 4, pp. 474–479, 2015.

[6] U. development Team, "UAVCAN Communication Protocol," 2014.

[7] Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Nayyar, A. (2020). Emerging use of UAV's: secure communication protocol issues and challenges. In Drones in smart-cities (pp. 37-55). Elsevier.

[8] Oudina, Z., Derdour, M., Dib, A., Bouhamed, M. M. (2024, April). Empirical Analysis of the Security Threats and Risks that Drones Face, Represent, and Mitigation. In 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-8). IEEE.

[9] Mekdad, Y., Acar, A., Aris, A., Fergougui, A. E., Conti, M., Lazzeretti, R., Uluagac, S. (2024). Exploring Jamming and Hijacking Attacks for Micro Aerial Drones. arXiv preprint arXiv:2403.03858.

[10] A. Shafique, A. Mehmood, and M. Elhadef, "Survey of security protocols and vulnerabilities in unmanned aerial vehicles," IEEE Access, vol. 9, pp. 46927–46948, 2021

[11] R. Gupta, A. Nair, S. Tanwar, and N. Kumar, "Blockchain-assisted secure UAV communication in 6G environment: Architecture, opportunities, and challenges," IET Commun., vol. 15, no. 10, pp. 1352–1367, 2021

[12] Yavuz, A. A., ALAGÖZ, F., Anarim, E. (2010). A new multi-tier adaptive military MANET security protocol using hybrid cryptography and signcryption. Turkish Journal of Electrical Engineering and Computer Sciences, 18(1), 1-22.

[13] A. Allouch, O. Cheikhrouhou, A. Koubâa, M. Khalgui,T. Abbes, "MAVSec: Securing the MAVLink Protocol for Ardupilot/PX4 Unmanned Aerial Systems," 2019 15th International Wireless Communications Mobile Computing Conference (IWCMC), pp. 621-628, 2019

[14] Koubâa, Anis, et al. "Micro Air Vehicle Link (MAVLink) in a Nutshell: A Survey." IEEE Access 7 (2019): 87658-87680

[15] C. Rani, H. Modares, R. Sriram, D. Mikulski, and F. L. Lewis, "Security of unmanned aerial vehicle systems against cyber-physical attacks," The Journal of Defense Modeling and Simulation., vol. 13, issue. 3, pp. 331-342, July. 2016

[16] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, "Inferring Internet denial-of-service activity," ACM Trans. on Computer Systems., vol. 24, issue. 2, pp. 115–139, May. 2006

[17] Vasconcelos, G.; Miani, R.S.; Guizilini, V.C.; Souza, J.R.Evaluation of dos attacks on commercial wi-fi-based uavs. Int. J.Commun. Netw. Inf. Secur. 2019, 11, 212–223

[18] Airlangga, G.; Liu, A. A Study of the Data Security Attack and DefensePattern in a Centralized UAV–Cloud Architecture. Drones 2023, 7,289. https://doi.org/10.3390/drones7050289

[19] J. Chen, Z. Feng, J.-Y. Wen, B. Liu, and L. Sha, "A container-based dos attack-resilient control framework for real-time uav systems," 2018.

[20] O. Westerlund and R. Asif, "Drone hacking with raspberry-pi 3 and wifi pineapple: Security and privacy threats for the internet-of-things," in 2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS), Feb 2019, pp. 1–10.

[21] G. Choudhary, V. Sharma, T. Gupta, J. Kim, I. You, Internet of drones(IoD): Threats, vulnerability, and security perspectives, in: The 3rdInternational Symposium on Mobile Internet Security, no. 37, 2018, pp.

[22] Hoang, T.M.; Nguyen, N.M.; Duong, T.Q. Detection ofeavesdropping attack in UAV-aided wireless systems: Unsupervisedlearning with one-class SVM and k-means clustering. IEEE Wirel.Commun. Lett. 2019, 9, 139–142

[23] Shivaramakrishna, D., Nagaratna, M. (2023). A novel hybrid cryptographic framework for secure data storage in cloud computing: Integrating AES-OTP and RSA with adaptive key management and Time-Limited access control. Alexandria Engineering Journal, 84, 275-284.

[24] Mandal, P., Roy, L. P., Das, S. K. (2023, June). Hybrid Cryptographic Technique of Data Security for UAV Applications. In 2023 3rd International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE.

[25] Madhu, A., Kumar, H. M. (2021, June). Positioning Optimization of Drones using IMU and Securing UAV Communication by implementing Hybrid Cryptosystem. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 681-686). IEEE.

[26] Oudina, Z., Derdour, M., Bouhamed, M. M. (2022, October). Testing cyber-physical production system: Test methods categorization and dataset. In 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-8). IEEE.

[27] Oudina, Z., Dib, A., Yakoubi, M. A., Derdour, M. (2024). Comprehensive Risk Classification and Mitigation in the Petroleum Cyber-Physical Systems of the Oil and Gas Industry. International Journal of Safety Security Engineering, 14(1).

[28] Oudina, Z., Derdour, M., Dib, A., Yaakoubi, M. A. (2024). Identifying and Addressing Trust Concerns in Cyber-Physical Systems for the Oil and Gas Industry. Ingénierie des Systèmes d'Information, 29(2).

[29] Oudina, Z., Derdour, M., Dib, A., Tachouche, A. M. A. (2023, October). Model Based System Engineering for trust SCADA and ICS Systems in Oil Gas Industry. In 2023 5th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-8). IEEE.

# VerChain: Blockchain Based Certificate Degree Attestation and Verification in Algeria

Rofaida Khemaissia[1,*,†] and Ala Djeddai[2,†]

[1] *Laboratory of Mathematics, Informatics and Systems (LAMIS), Echahid Cheikh Larbi Tebessi University, Tebessa,12002, Algeria*

[2] *Laboratory of Computer Science and Applied Mathematics (LCSAM), Chadli Bendjedid El-Tarf University, B.P 73, El Tarf 36000, Algeria*

**Abstract**

Many companies from all over the world are prone to counterfeit academic certificates, which would trigger colossal material losses, and it is time-consuming for universities to undertake this process by spending a huge number of budgets annually. In this case, the need for a degree verification process becomes important, as a famous platform Block-cert issued by Massachusetts Institute of Technology, aims to facilitate degree verification between the university and the company, empowered with the adoption of blockchain technology as a proof of existence by leveraging the immutability and availability features. In the academic and industrial community, several degree verification platforms have been proposed and their feasibility proved; however, in this paper, we propose the VerChain Algerian degree verification platform as a vital step towards fighting degree forgery and digitalization, which calls for the interference of several government ministries to render the process realistic and practical. VerChain is a blockchain based system for certified degree attestation and verification that aims to maintain security and data privacy, where it is managed by several Algerian ministries in which everyone has its responsibilities and access restrictions.

**Keywords**

Blockchain, Hyperledger Fabric, Data Integrity, Smart Contracts, Privacy, Certificate attestation and verification.

## 1. Introduction

Under the supervision of the Algerian Minister of High Education and Scientific Research, 377,000 Algerian students graduate annually from public or private universities. Due to this huge number, enormous budgets are spent by the Algerian government to verify the validity and existence of a student's degree, let alone the wasted time beyond the process. Digitalization economizes costs and renders the process more trustworthy and rapid.

Blockchain is a well-known technology that has become widespread because of several concerns related to a trustless environment. According to the level of permission blockchain can be seen as two main categories, permissionless and permissioned, regarding the access rights to the network and the level of centralization, permissionless supports a wide public network in which every participant can partake and join the network as a blockchain peer and acts as a miner. Otherwise, a permissioned blockchain is known as private ledger, in which access to the network is under control and restricted only to legal nodes; it is known as semi-centralized.

[*] Corresponding author.

[†] These authors contributed equally.

🖂 khemaissia.rofaida@univ-tebessa.dz (R. Khemaissia) ;a.djeddai@univ-eltarf.dz (A. Djeddai) ;

iD 0000-0002-4092-537X (R. Khemaissia); 0000-0002-9354-4108 (A. Djeddai)

By leveraging different blockchain 'merits such as immutability that helps to remain unalterable information, besides to auditability, availability and persistency. By adopting this technology, numerous academic and industrial solutions have been introduced for lucrative and free use. As centralizing the verification process is not supported by the Algerian government, the decentralization feature of the distributed ledger blockchain can handle a bulk of issues.

The research introduces VerChain, an Algerian framework for verifying degree certificates. This solution implements a system of separated roles to minimize potential errors and enhance both security and privacy measures. VerChain consists of three main actors starting with the ministries; then, the certificate issuer (i.e., University) and certificate verifier (national company), all of which are managed by blockchain, which acts as a mediator between actors. Because Blockchain maintains lightweight information for future scalability issues, VerChain replaced the certificate degree hard copy with a one-way generated hash as a proof of existence (SHA 256 algorithm). The hash code is generated by mashing several pieces of information to garner a particular hash that represents a unique identifier (ID) of a given certificate, thereby preserving the degree integrity and privacy of the student credentials. Blockchain technology is adopted as a distributed and decentralized database; however, altering (write, update, or delete) with saved information is more likely impossible because the information is replicated across a peer-to-peer network, which would help to increase availability and accessibility at any request. Since the National Authority for the Protection of Personal Data (ANPDP) [1] imposes that whatever the operators must undergo the privacy rules, students' credentials are provided privately by the Ministry of the Interior, Local Authorities, and Regional Planning, and the certificate verifier must undergo the authority rules by using the required information that the law allows.

The rest of the paper covers related works in Section 2, where Section 3 provides details about VerChain architecture and its main components, and Section 3 presents an example of the VerChain scenario executed using smart contracts. A possible implementation is presented in Section 4. Finally, Section 5 summarizes the paper.

## 2. Related Works

Digitalization is considered one step ahead for quick and better information processing, and it prevails in different fields, encompassing distance learning that grants students the opportunity to garner an online diploma. From here, the rate of degree or diploma falsification is increasing, that would be the reason behind adopting blockchain to create a trustworthy environment by many research, since The Massachusetts Institute of Technology (MIT) has put forward BlockCerts [2] an open and freely standard for academic certificate verification that was implemented on bitcoin as a public ledger, the system boosts a decentralized and trust student credentials verification via sending invitations to whom willing to join the system, through creating a student accounts then share it with the corresponding university/institute, where issuing only the degree onto the BlockCerts, in order to keep the credentials integrity BlockCerts through substituting the genuine degree hard copy with a generated irreversible hash code by employing a hashing algorithm. Another Blockchain solution, Docschain [3], was introduced to treat bockcerts' limitation platform under the consortium Hyperledger fabric. In addition to employing the hash function, Docschain utilized optical use character recognition (OCR) to extract the original data from the degree hard copies. Furthermore, Docschain integrated Internet of Things (IoT) devices using an IoT camera for executing the read operation. The CVSS [4] is another certificate verification platform implemented under the Ethereum Blockchain in Vietnam, where its functionalities are issuing, verifying, and retrieving student degrees.

In addition, researchers have proposed several blockchain-based degree issuing and verification approaches under the implementation of various blockchain platforms such as bitcoin [6,7], Ethereum [8,9], Hyperledger Fabric [10,11], multichain [12], [13], and tangle [14]. For more information about other works on the paper topic, the reader can refer to the systematic review [15] on integrating blockchain as a trust-distributed database for academic certificate verification.

In Algeria, great efforts have been made by the startup Takawen, which has attempted to fight forged training certificates using the Algerian certification and verification portal [5]. However, this experience does not apply in university degree certification, and the portal has solely settled for employing a traditional database system (Create/Read/Update/Delete), which is prone to attacks such as SQL-injection attacks, weak authentication attacks, and potentially privilege abuse. In this situation, blockchain presents an ideal solution to overcome these issues, although deploying blockchain across the entire national territory may face new challenges.

## 3. VerChain Architecture and its main Functionalities

Figure 1 depicts the VerChain components and their associated actors, showcasing all necessary smart contracts. The blockchain smart contracts proposed by VerChain enable users to engage with the BC network. The subsequent section offers comprehensive explanations of VerChain components and their authorized users, as well as their interactions.

### 3.1. Users Roles and Responsibilities

This section outlines the various import roles within VerChain. By breaking down these roles, the system enhances its security measures and clearly defines the responsibilities of each participant. This separate role contributes to the overall integrity and accountability of the system.

**Ministry of the Interior, Local Authorities and Regional Planning (Algeria):** Its main role is to manage blockchain data on persons who are under degree preparation. Each individual must be enrolled using the national identification number, which serves as a unique identifier. This distinct number can be utilized to connect every person to their blockchain-stored certificate. Due to the sensitive nature of citizen data, MILARP is the sole entity authorized for this task. MILARP has the capability to oversee access controls and identity management based exclusively on personal information.

**Minister of Higher Education and Scientific Research (Algeria):** Its main role is to register new certificate issuers, such as universities and institutes, because it is the main authority with this right. Every certificate issuer is registered using a unique ID along with its specific information. MHESR can manage access controls and identity management based only on certificate issuer data.

**Minister of Vocational Training and Education MVTE (Algeria):** It performs the same function as the MHESR, but lacks the authority to register certificate issuers in the fields of higher education and scientific research.

**Certificate Issuer:** This is the only authority that delivers the certificate. Its role is attributed to the Ministry of Education and Higher Education, or MVTE, by providing a blockchain certificate. It uses this certificate and the Algeria ID of a person to store the certificate degree in the blockchain.

**Certificate verifier:** This can be any organization that can verify the certificate degree of persons. For example, a company could verify the certificate degree of every candidate registered for a job.

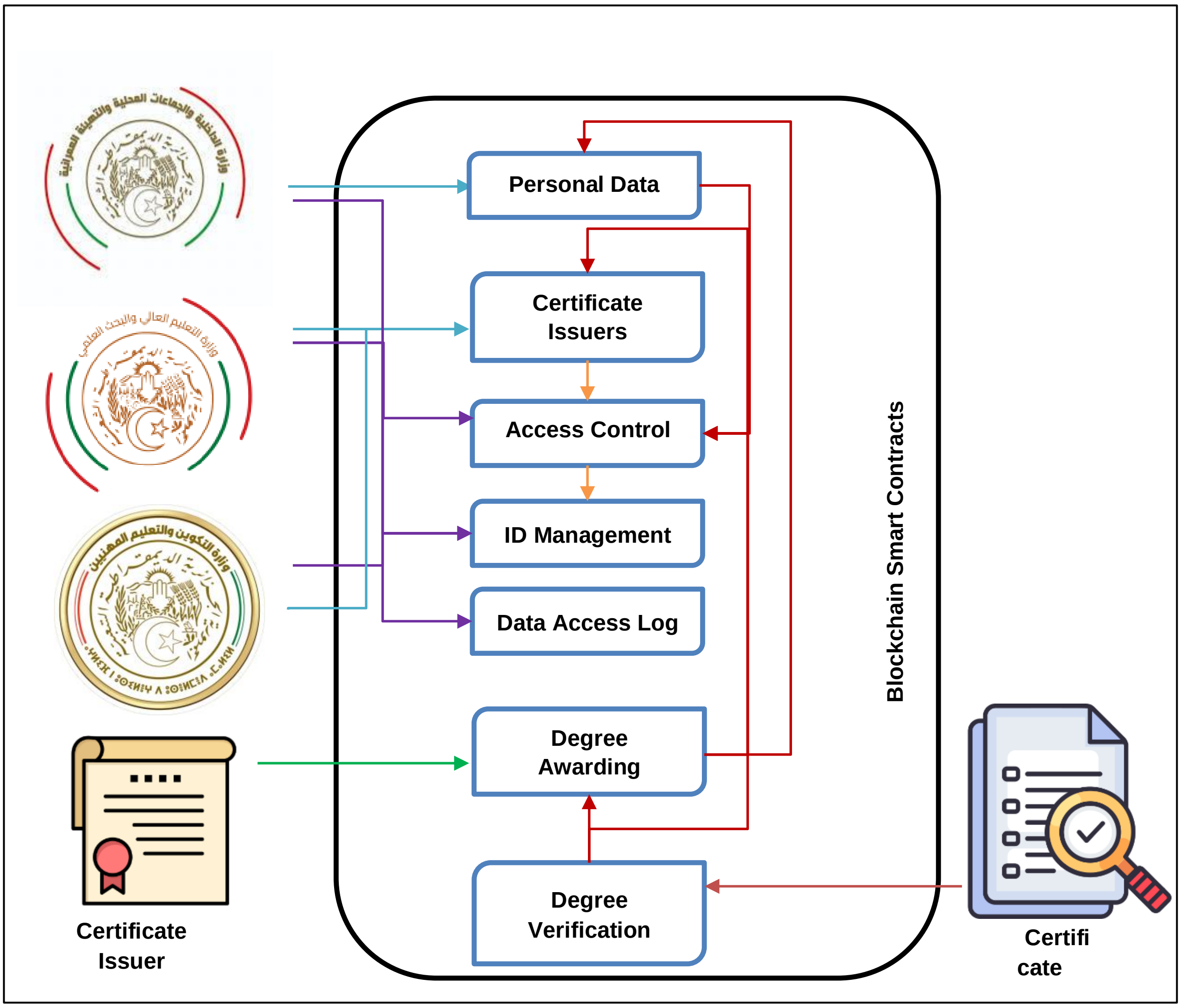


**Figure 1:** The Main Components of VerChain Architecture along with theirs Interactions

### 3.2. Blockchain Components

In this section, we provide descriptions of VerChain components where everyone is associated 0with one or several roles. Every component is considered to be a blockchain smart contract that ensures the services offered by that component. All components must interact with Access Control to verify whether the user has access authorization to use the target components.

**Personal data:** This manages data related to persons who have received certificate degrees from certificate issuers. These data are considered sensitive; therefore, we proposed that they be controlled and managed by the Minister of the Interior. Every person is registered with a unique Algeria ID.

**Certificate Issuers:** It uses MHESR or MVTE to control and manage the data related to certificate issuers. The MHESR or MVTE stores every registered issuer using its unique ID along with its information. Therefore, every unregistered issuer is not allowed to deliver a certificate. The MHESR or MVTE stores authorized degrees that can be delivered to every issuer.

**Identity Management**: It has two main tasks: creating identities and registering new VerChain users such as certificate issuers and verifiers, and verifying whether a given identity is valid using the blockchain. The IDM component returns a registration certificate for every accepted registration demand, which contains critical information about enrolling VerChain users with different roles. All identity information is stored in the BC to protect VerChain from fraudulent identities. Only MILARP, MVTE, and MHESR interacted with IDM.

**Access Control:** Access controls are defined for VerChain components, which are only used by MILARP, MHESR, and MVTE. For example, the BC administrator places restrictions on accessing the verifier component of the certificate. The restriction can be, for example, to provide access for a period of time to some verifiers.

**Access Log:** This stores all the operations performed by VerChain users. The main objective is to perform an advanced verification that detects inconsistencies in the history of authorization and access control components. It merges the data from all ledgers and performs advanced checks. For example, if a certificate issuer changes the data about a delivered certificate degree, this operation can be checked using audit verification. The auditing process is initiated by MILARP, MHESR, or MVTE.

**Degree Awarding:** This is invoked by a legitimate certificate issuer to register a new degree. In this situation, the certificate issuer and person ID must already be stored with their information using the certificate issuer data and personal data components, respectively. This restriction ensures that legal issuers deliver certificates to legal persons. Every certificate must be stored with a unique ID that can be, for example, a hash calculated using the certificate data.

**Degree Verification:** A legitimate certificate verifier uses this component to check whether a given certificate degree has been delivered to a person. The verification process uses information from the degree award to complete the check using the data provided by the certificate, such as the certificate hash and person ID.

## 4. Example of Scenario in VerChain and their Smart Contracts

In this section, we present an example of a scenario that can be controlled using VerChain smart contracts. Publishing and verifying a given certificate on the blockchain. ElTarf University takes the role of a certificate issuer, whereas Oracle Company plays a certificate verifier role. Mohamed an Algerian Citizen takes the role of a candidate requesting the Oracle Job. The following steps explain how this scenario can be executed.

1. Oracle requires candidates for a given job of a Data Scientist. Therefore, it requires a master's degree in computer science.

2. The MILARP registers Mohamed on the blockchain using his national ID along with his required personal data such as first name, last name, birthdate, address, etc.

3. After receiving a certificate of registration from the MHESR, ElTarf University issued a Master of Computer Science to Mohamed Using his national ID and his Personal Data. VerChain then checks whether Algerian University has the right to this kind of degree using onchain data on universities.

4. After registering Mohamed for the Oracle job using the hash of his certificate degree and national ID, Oracle uses its certificate automatically given by the VerChain administrator to verify the given degree through blockchain smart contracts.

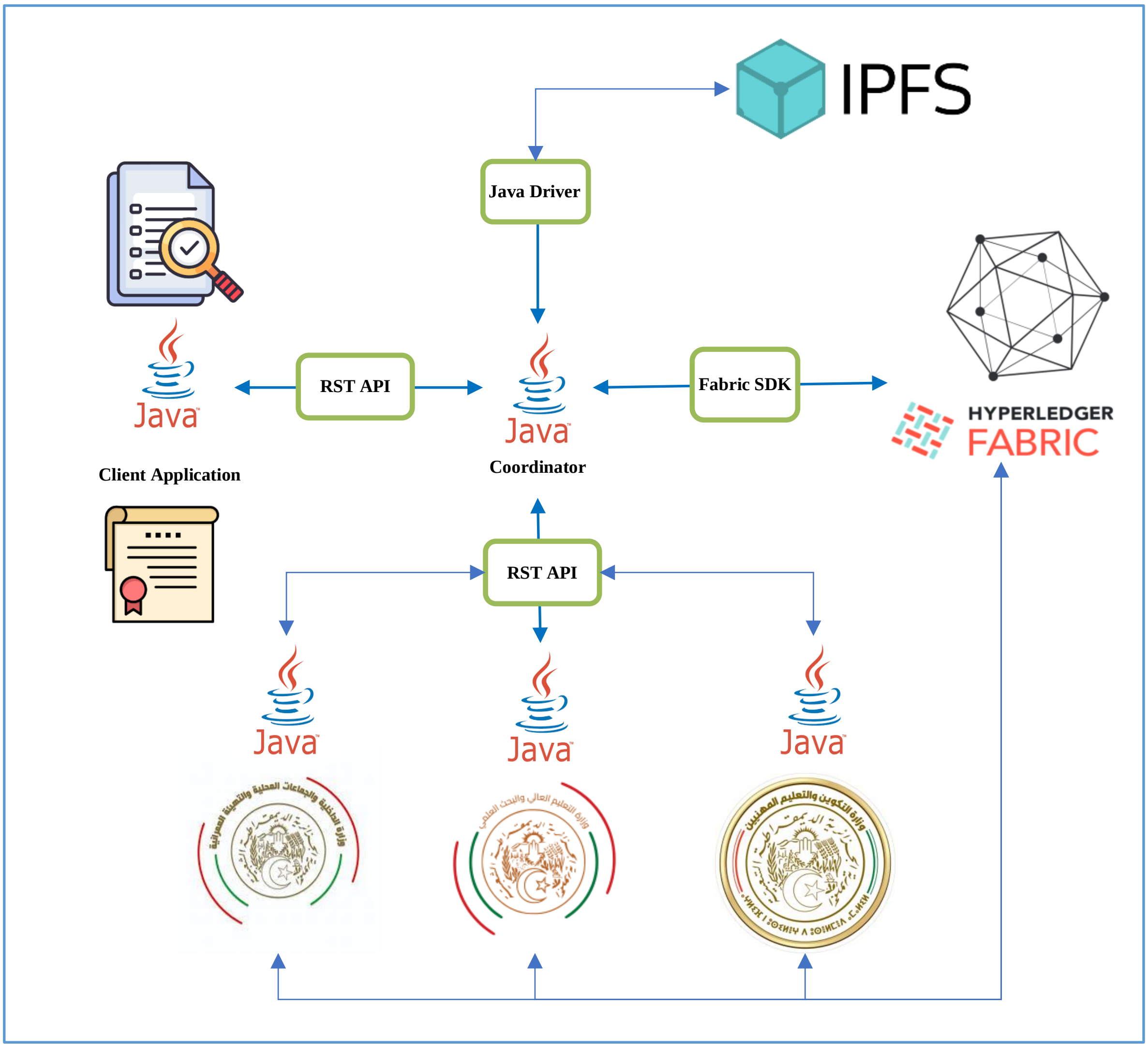


**Figure 2.** Implementation Architecture of the proposed VerChain.

## 5. Implementation

The proposed approach was implemented under Eclipse using various Java APIs, such as JSON and Fabric SDK. The implementation architecture is illustrated in Figure 2, where the main components are as follows:

*The Fabric Hyperledger Blockchain*[2] is used with the configuration of two organizations and one peer node for each one. The blockchain network uses CouchDB as a world state database and an ordering service. It uses a certificate authority for each organization. Six channels were created for identity management, degree warding, degree verification, Personal Data, Access Control, and logs. Six Fabric smart contracts were deployed using the Go language (one for each channel). The Hyperledger Fabric network used in VerChain is shown in Figure 3 where every channel is associated with its ledger and smart contract. Figure 4 illustrates the steps involved in executing a transaction in the fabric network for the chain code associated with the identity management channel.

[2] https://hyperledger-fabric.readthedocs.io/en/release-2.2/

*IPFS*[3]: The Interplanetary File System is a distributed file storage, where every file added to the IPFS is given a unique address derived from a hash of the file's content.

*Client Application*: a component that needs query results.

All these latter are interacted with the main program using their specific Java API.

### 5.1. Fabric Chaincodes and Distributed Ledgers in VerChain

Every peer in HLF has its local database (ledger), which contains all transactions executed by the network via HLF chaincodes. Thus, each peer can have several installed chain codes for a single HLF channel. The distributed ledgers in the HLF are updated by executing smart contracts with blockchain external users. Our work proposes the use of six distributed ledgers (one for every VerChain component), where each ledger is associated with one chaincode and several peers. These ledgers store critical data on VerChain operations, such as identity management degree awarding, degree verification, personal data, and operation logs.

#### 5.1.1. Identity Management

The IDM chaincode defines functions executed by HLF peers for managing the participants registered by the BC, such as certificate issues (e.g., universities), certificate verifiers (e.g., enterprises), and organizations Algerian Ministers. This chaincode is installed on a channel identified by the same name, Identity Management," and is associated with a local ledger that saves information about registered entities. The IDM chaincode uses the Golang structure, which is illustrated by *listing 1* whereas the IDM ledger uses the JSON key-value format given by *listing 2* as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the IDM ledger.

```
type Degree Issuer struct {
    DIsId string `json:"DISId"`
    DIsName string `json:"DIsName"`
    DIsType string `json:"DIsType"`
    DIsAddress string `json:"SIsAddress"`
}
type Organization struct {
    OrgId string `json:"OrgId"`
    OrgName string `json:"OrgName"`
    OrgType string `json:"OrgType"`
    OrgAddress string `json:"OrgAddress"`
}
```

**Listing 1.** The Golang Structures used by the Identity management Chaincode

[3] https://ipfs.io/

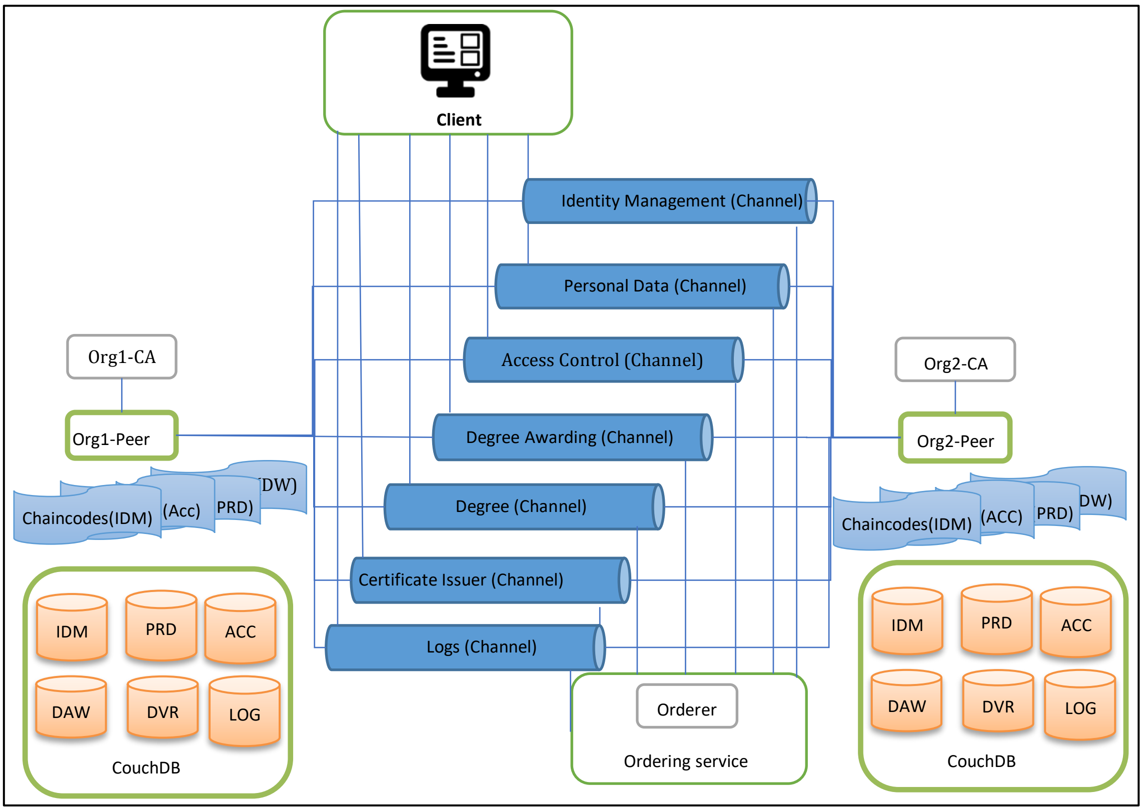


**Figure 3.** The Hyperledger Fabric Network Used by The Proposed Method

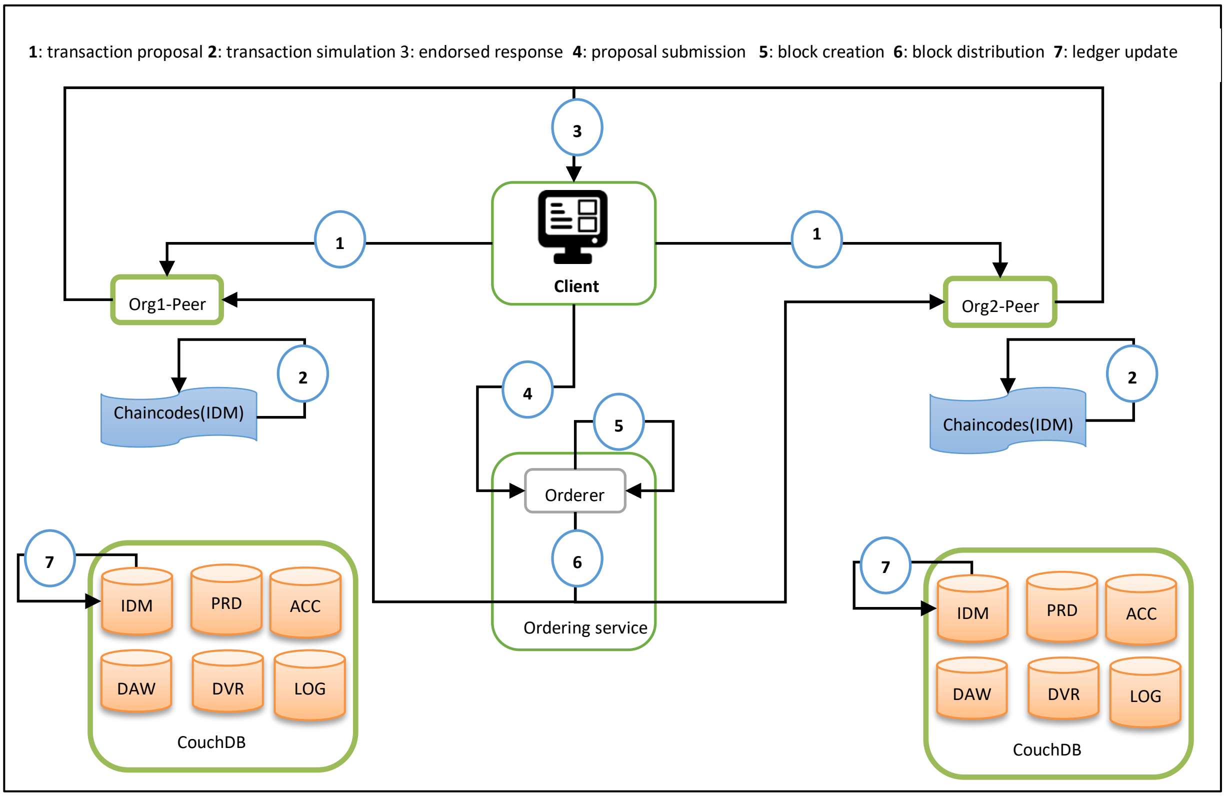

1: transaction proposal 2: transaction simulation 3: endorsed response 4: proposal submission 5: block creation 6: block distribution 7: ledger update
3
Client
1
1
Org1-Peer
Org2-Peer
2
Chaincodes(IDM)
2
Chaincodes(IDM)
4
5
Orderer
6
Ordering service
7
IDM
PRD
ACC
DAW
DVR
LOG
CouchDB
7
IDM
PRD
ACC
DAW
DVR
LOG
CouchDB

**Figure 4.** Fabric Steps to Execute a New Transaction of the Chain-code Associated with Identity Management Channel

```
{"DIsID": "a21khjh44, ” ”","DIsName":"National Training Center", "DIsAddress":"ElTarf",
"DIsType":"National Institution"}, {"OrgID": "a21kh32fr","OrgName":"ElTarf University",
"OrgAddress":"ElTarf", "OrgType":"University”}
```

**Listing 2.** Example of JSON key value Structure used by Identity management chaincode

| Function | Description | Restricted Access |
|---|---|---|
| CreateDegreeIssuer | Create new degree issuer using the description given by the invocation parameters | MHESR, MVTE |
| CreateDegreeVerifier | Register new degree verifier using the description given by the invocation parameters | MHESR, MVTE |
| GetAllCIs | Get all certificate issuers that are already registered by the BC | MHESR, MVTE |
| updateCI | Updating existed certificate issuer with new information. | MHESR, MVTE |

**Table 1.** Some Smart Contract Functions that are Implemented by identity management Chaincode

### 5.1.2. Access Control chaincode

The AC chain code defines the functions executed by HLF peers for managing access controls associated with registered participants. This chaincode is installed on a channel identified by the same name, Access Control,” and is associated with a local ledger that saves information about data access control and policies. The AC chaincode uses the Golang structure, which is illustrated by *listing 3*, whereas the AC ledger uses the JSON key-value format given by *listing 4* as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the AC ledger.

```
type Access Control struct {
        RequesterID string `json:" RequesterID "`
        Permissions [] PermissionType `json:" Permissions"`
        }
type PermissionType struct {
        Permission string `json:"Permission”`
        PermissionType string `json:" PermissionType"`}
```

**Listing 3.** An example of a Golang Structure used by the Access control Chaincode

```
{"RequesterID":"a21kh32fr", "Permissions":[{"Permission":"Issuing degree certificate",
"PermissionType":"Write":[{"Permission": "Verifying degree certificate",
"PermissionType":"Read"}]}
```

**Listing 4.** Example of JSON key value Structure used by the Access control Ledger

| Function | Description | Restricted Access |
|---|---|---|
| CreateACC | Create new access control for a given requester | MILARP, MHESR, MVTE |
| UpdateACC | Update a given access control for a given requester. | MILARP, MHESR, MVTE |
| GetACC | Get the access control associated with a given requester. | MILARP, MHESR, MVTE |
| RemoveACC | Remove an access control associated with a given requester. | MILARP, MHESR, MVTE |

**Table 2.** Some Smart Contract Functions that are Implemented by the Access control Chaincode

### 5.1.3. Personal data Chaincode

The PSA chaincode defines functions executed by BC peers for managing saved data about persons who have been awarded degrees. This chaincode is installed on a channel identified by the same name “Personal Data” and it is associated with a local ledger that saves information about awarding people which can be queried by members given by table 3. The PSA chaincode uses the Golang structure, which is illustrated by listing 5, whereas the SC ledger uses the JSON key-value format given by listing 6 as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the SC ledger.

```
type personal data struct {
        StudentNIN string `json:" StudentNIN"`
        StudentFirstName string `json:" StudentFirstName "`
        StudentLastName string `json:" StudentLastName "`
        StudentBirthDate string `json:" StudentBirthDate"`
        StudentEmail string `json:"StudentEmail "`
        StudentNumT string `json:"StudentNumT "`
        StudentAdr string `json:"StudentAdr"`
}
```

**Listing 5.** The Golang Structure used by the Personal Data Chaincode

```
{"StudentNIN ": "1099992515483625",
 "StudentLastName": "Mohammed",
 "StudentLastName": "Saadi",
  "StudentBirthDate": "12/9/2000",
 "StudentEmail": " Mohammed.Saadi@univ-elTaref.dz",
 "StudentNumT": "0655251649",
 " StudentAdr": "ElTarf, Algeria",
}
```

**Listing 6.** Example of JSON key value Structure used by the Personal Data Ledger

| Function | Description | Restricted Access |
|---|---|---|
| CreatePrData | Create new personal data on the blockchain. | MILARP |
| GetPrsData | Retrieve from the BC the personal data associated with a given NIN | MILARP |
| UpdatePrsData | Update an existing personal data for a given NIN | MILARP |

**Table 3.** Some Smart Contracts Functions that are Implemented by the Personal Data Chaincode

```
type Degree Awarding struct {
        StudentNIN string `json:" StudentNIN"`
        DegreeID string `json:"DegreeId"`
        DegreeDeliverID string `json:"DegreeDeliverID"`
        DegreeTitle string `json:" DegreeTitle"`
        DegreeType string `json:" DegreeType"`
        DegreeSpeciality string `json:" DegreeSpecialty"`
        DegreeMention string `json:" DegreeMention"`
        SignedBy [] SignatureType `json:"SignedBy"`
type SignatureType struct {
        PersonNIN string `json:" PersonNIN”`
        PersonQuality string `json:" PersonQuality"`}
}
```

**Listing 7.** The Golang Structure used by the Service Composition Chaincode

```
{"StudentNIN ": "1099992515483625",
 "DegreeDeliverID": "a21kh32fr",
 "DegreeID": " as1kh2562",
  "DegreeTitle": "Certificate issuing and verification using Blockchain",
 "DegreeType": "Master of Engineering",
 "DegreeSpeciality": "Information Security",
 "DegreeMention": "Very Good",
 "SignatureType":[
   {“PersonNIN”: "1099260215022518”,
   “PersonQuality”: "Dean of the faculty of Sciences and Technologies”},
   {“PersonNIN”: "1099512502160217”,
   “PersonQuality”: "Rector of Sciences and Technologies”}
]}
```

**Listing 8.** Example of JSON key value Structure used by the Service Composition Ledger

### 5.1.4. Degree Awarding Chaincode

The DAW chaincode defines functions executed by BC peers for managing saved data about degree- awarding processes, such as degree delivery institutions and target students. This chaincode is installed on a channel identified by the same name "Degree Awarding" and it is associated with a local ledger that saves information about the degree awarding process which can be queried by members given by table 4. The SC chaincode uses the Golang structure, which is illustrated by listing 7, whereas the SC ledger uses the JSON key-value format given by listing 8 as a representation of the same GO structure. The Go functions use marshal and unmarshal methods to manipulate the JSON strings and store them in the SC ledger.

| Function | Description | Restricted Access |
|---|---|---|
| GetDegreeAW | Get a degree data using its DegreeID. | MHESR, MVTE and theirs registered degree awarding organizations. |
| CreateDegreeAW | Register a new degree awarding data. | Registered degree awarding organizations |
| UpdateDegreeAW | Update an existing degree data | Registered degree awarding organizations |

**Table 4.** Some Smart Contracts Functions that are Implemented by the Service Composition Chaincode

## 6. Evaluation

### 6.1. Experiment Configuration

To validate the functionality and test the performance of our approach, several experiments were performed. Experiments were performed on a machine with an Intel Core i7 processor running with a 1.8 GHz clock speed, 16 GB memory, 128 GB SSD, and 1 TB for storage. The components of the fabric network were deployed as Docker 2.3 images (organizations, certificate authorities, peers, CouchDB. etc.). With regard to the implementation architecture, the coordinator is implemented as a JAVA REST application that uses Tomcat 9 as the resource server. All service ministries and clients (certificate issuers and verifiers) are depicted as JAVA applications that communicate with servers using the REST API. Our implementation uses several Java API in different processes, such as the REST API, IPFS API, and Fabric SDK.

## 7. Conclusion

This paper proposes VerChain, an Algerian platform blockchain-based system for the generation and verification of academic degrees and student credentials. First, the proposal presents the existing solutions of several blockchain implementations that share similar endeavors. Subsequently, it sheds light on Algerian architecture that detailed the interaction between certificate verifiers and government ministries, along with their role and access restrictions. Moreover, the study underlined the privacy of data related to students' credentials, issuers, and even certificates using access control. Therefore, verchain-managed access is allowed only by a valid identity and a specific role. VerChain ensures automatic certificate issuing and verification, which limits errors and omits manual processes. Traceability is guaranteed through the log component to start the auditing process to detect any error or illegal access, and the communication between clients and VerChain uses TLS for better security and privacy preservation.

## References

[1] ANPDP: https://anpdp.dz/fr/accueil/.

[2] P. Schmidt. (Oct. 24, 2016). Blockcerts—An Open Infrastructure for Academic Credentials on the Blockchain MLLearning, MIT Media Lab. [Online] Availble: https://medium.com/mit-media-lab/blockcertsan- open-infrastructure-for-academic-credentials-on-the-blockchain-899a6b880b2f

[3] S. Rasool, A. Saleem, M. Iqbal, T. Dagiuklas, S. Mumtaz and Z. u. Qayyum, "DocsChain: blockchain-Based IoT Solution for Verification of Degree Documents", IEEE Transactions on Computational Social Systems, vol. 7, no. 3, pp. 827-837, June 2020.

[4] Nguyen, D. H., Nguyen-Duc, D. N., Huynh-Tuong, N., & Pham, H. A. (2018, December). CVSS: a blockchainized certificate verifying support system. In Proceedings of the 9th international symposium on information and communication technology (pp. 436-442).

[5] Algerian Certification & Verification Portal access: https://www.takawen.dz/about-us-2/.

[6] M. Sharples and J. Domingue, "The blockchain and kudos: A distributed system for educational record, reputation and reward," in Proc. Eur. Conf. Technol. Enhanced Learn. Cham, Switzerland: Springer, 2016, pp. 490–496.

[7] A. S. de Pedro Crespo and L. I. C. García, "Stampery Blockchain Timestamping Architecture (BTA)–Version 6," 2017, arXiv:1711.04709. [Online]. Available: http://arxiv.org/abs/1711.04709

[8] T. Kanan, A. T. Obaidat, and M. Al-Lahham, "SmartCert blockchain imperative for educational certificates," in Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT), Apr. 2019, pp. 629–633.

[9] B. Boeser. (Jul. 2017). Meet Truerec by Sap: Trusted Digital Credentials Powered by Blockchain | Sap News Center. Accessed: Jan. 29, 2020. [Online]. Available: https://news.sap.com/2017/07/meet-truerec-by-saptrusted-digital-credentials-powered-by-blockchain/

[10] "A permissioned blockchain-based system for verification of academic records," in Proc. 10th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS), Jun. 2019, pp. 1–5.

[11] J. A. Otuya, "A blockchain approach for detecting counterfeit academic certificates in Kenya," M.S. thesis, Dept. Inf. Technol., Strathmore Univ., Nairobi, Kenya, 2019.

[12] G.-A. Dima, A.-G. Jitariu, C. Pisa, and G. Bianchi, "Scholarium: Supporting identity claims through a permissioned blockchain," in Proc. IEEE 4th Int. Forum Res. Technol. Soc. Ind. (RTSI), Sep. 2018, pp. 1–6.

[13] R. Arenas and P. Fernandez, “CredenceLedger: A permissioned blockchain for verifiable academic credentials,” in Proc. IEEE Int. Conf. Eng., Technol. Innov. (ICE/ITMC), Jun. 2018, pp. 1–6.

[14] A. Wahab, M. Barlas, and W. Mahmood, “Zenith certifier: A framework to authenticate academic verifications using tangle,” J. Softw. Syst. Develop., vol. 2018, no. 370695, p. 13, 2018.

[15] A. Rustemi, F. Dalipi, V. Atanasovski and A. Risteski, "A Systematic Literature Review on Blockchain-Based Systems for Academic Certificate Verification," in *IEEE Access*, vol. 11, pp. 64679-64696, 2023, doi: 10.1109/ACCESS.2023.3289598.