Enhancing Card Fraud Detection Using Large Language Model (LLM)

SUMIT[1], AMANDEEP[2], DIPTI [3],ARJOO [4]
M.Sc. Computer Science [1, 3, 4], Artificial Intelligence & Data Science, GJUS&T HISAR
Assistant Professor[2], Artificial Intelligence & Data Science, GJUS&T HISAR
sumetjangra36@gmail.com

Abstract- While digital financial transactions have become more convenient, card fraud threats have also risen — especially in card-not-present (CNP) and identity theft cases. The conventional fraud detection mechanisms are unable to tackle the shifting nature of fraud over a period of time, data impairment, and limitation in transparency. We investigate the possibility of applying Large Language Models (LLMs), like GPT, to predictive models of fraud built on structured transaction data that is represented as unstructured natural language for both enhanced detection capability and interpretability. With a combination of LLMs and models such as Logistic Regression and XGBoost, the hybrid system provides higher detection accuracy with human-readable explanations. The paper shows that the generalizing and adaptive nature of LLMs enables them to improve fraud detection systems that comply with regulatory requirements.

Keywords: Card Fraud Detection, Large Language Models (LLMs), Explainable Artificial Intelligence (XAI), SMOTE-Tomek, Natural Language Processing (NLP), XGBoost, Logistic Regression**.**

I. INTRODUCTION

The battle against financial fraud is growing in this digital era, thanks to the proliferation of online banking, digital wallets and mobile financial services. Fraud in credit cards, in particular, is a persistant and cat-and-mouse game, costing billions of dollars in annual losses worldwide. With more than $5.8 billion in financial fraud losses based on U.S. Federal Trade Commission estimates last year[1], the need for robust prevention systems is more pressing than ever. Fraud comes in many shapes identity theft, phishing, card-not-present fraud, application fraud, account fraud, and account takeover – and each includes its unique set of challenges for detection[1].

The majority of the known fraud detection systems are rule based and statistical models that model the suspicious activities with some predefined thresholds/behavior and also some historical pattern matching. Although these approaches have provided a degree of success in detecting known fraud profiles, they tend to have high false positive rates[2], lack scalability, and lack adaptivity to new or emerging fraud profiles. Moreover, these systems falter when it comes to handling the major class imbalance in datasets of fraud, as only a small percentage—usually below 1% of all the transactions are fraudulent ones.

Thanks to machine learning (ML), there is now new potential in fraud detection wherein the system can learn from the past and get better over time. Traditional ML models (e.g., decision trees, logistic regression, and ensemble methods such as Random Forest or XGBoost) have been shown to outperform rule-based systems[2][3]. That said, they cannot be generally applied to high dimensional and unstructured distorted inputs, since they still require carefully designed features, and are designed only for structured data.

More recently, Large Language Models (LLMs) have made a breakthrough in the artificial intelligence community by showing impressive performance on comprehension and generation of human-like text. These models, first developed for NLP problems, have been successful for fraud detection, analyzing transaction descriptions, behavioral logs, and customer communication [3][4].LLMs can encode fine-grained semantic associations and contextual clues that could be indicative of fraudulent intent or behavior.

This work investigates a hybrid fraud detection model that interfaces both the context-rich comprehension of LLMs and the structured learning functionalities of standard ML models and anomaly detectors. The combination of LLMs with models such as Logistic Regression, XGBoost, Autoencoders and Isolation Forests allows the approach to provide high accuracy, few false positives, and improved transparency. First Experiments and Validation The dataset for the experiments is the UCI Credit Card Fraud Detection[5]. This article offers perspectives on modeling, performance measurement, [9] real-time application and future prospects of AI-agile model for fraud detection.

II.LITREATURE REVIEW

Historically, FDSs have been based on rule-based systems, statistical methods (e.g., logistic regression and anomaly detection), and machine learning (ML) models, such as SVM, decision trees, and neural networks Although the above methods are effective in terms of pattern recognitions, they turn out to be less flexible, have high false positive rates, and show a wrong prediction performance behavior when the fraud strategies evolve [4], [5], [6]. Abdallah et al. Fraud detection methods have been studied across five domains such as credit cards and telecommunications [2]. They mentioned the major challenges in it like concept drift, class-imbalance, and requirement of real-time detection. This work highlighted the role of hybridization and flexible frameworks. Akash et al. Integration of statistical approaches with ensemble learning (e.g., XGBoost) to improve the sensitivity and specificity for their detection was recommended by [3]. They emphasized the need to retrain models frequently in changing environments. The rise of Large Language Models (LLMs) is creating a new paradigm in Fraud. Chkirbene et al. A survey covering industrial applications of LLMs for various sectors including finance and healthcare is offered in [10]. Computational cost, data privacy, and a lack of explainability were among the discussed challenges that the service could overcome by using domain-specific LLMs and LLM-as-a-Service (LLMaaS), they said. In [5], we developed a real-time phone fraud detection model based on combining LLMs with Automatic Speech Recognition (ASR). By analyzing call content, their system outperformed traditional metadata-based models in resisting spoofing, achieving 90.4% accuracy and 91.2% F1-score. Another recent work, AI-based LLM for Credit Card Fraud Detection [6] [13] provided a hybrid method of using prompt engineered LLMs using Regression and other traditional classifiers (Logistic Regression and XGBoost). It provided more interpretability for human auditors, serving as a middle ground between black-box models and the demands of real-world financial auditing.

III.METHODOLOGY AND MODEL DEVELOPMENT

This section describes a complete system for building a practical secure and efficient card fraud detection system. Through the combination of classic ML, Anomaly Detection, and LLMs, our aim is to guarantee high level of accuracy, Generalizibility and interpretability in both, offline and real-time scenarios.A detailed method to establish an effective card fraud detection system is described in this section. It consists of a number of important steps such as data collection, preprocessing, normalization, managing imbalanced class, splitting, model building, hyperparameter tuning, evaluation, and deployment. It is ensured that each phase has been developed in such a way that, the fraud detection model remains efficient and robust.

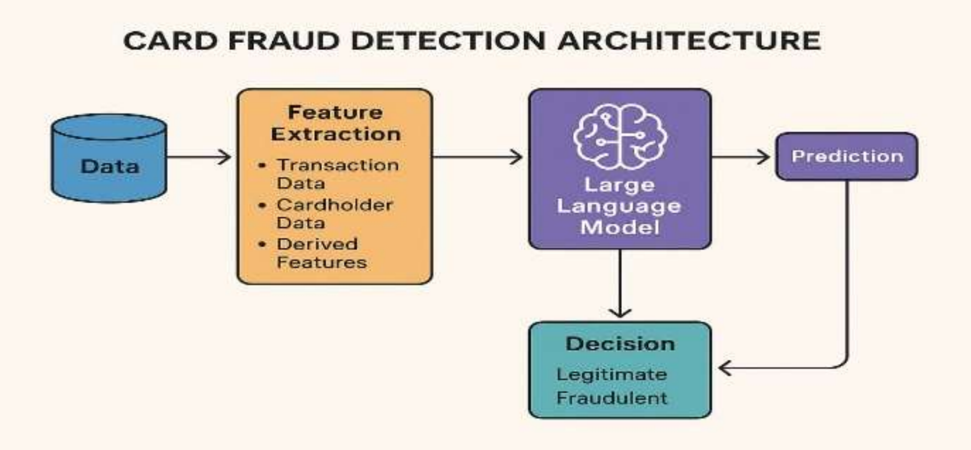


Fig. no 3.1 Architecture of the Model

A. Data Collection

We employ the UCI Credit Card Fraud Detection dataset[5] [14], which consists of 284,807 transactions recorded in two days, of which 492 are fraudulent transactions.
Features are primarily anonymized numerical features, derived through Principal Component Analysis (PCA), which transformed the original features in a way that preserved useful properties while hiding values.
Other raw features (e.g., Time, Amount) give a time and money context, which is essential to tell normality from anomalous behavior.

B. Data Preprocessing

- **Data Cleaning:** Elimination of redundant and incorrect records to make the dataset accurate. Preventing and removing outlying data that can corrupt model training.

- **Feature Engineering :** Extracting and generating features of importance such as : Transaction time features (hour of day, weekday/weekend). Aggregated User Behaviour measures(mean transaction amount, no of transactions).

- **Missing Value Handling:** Since we have no missing data this time, typically missing values would be imputed by mean, median or by using predictive modeling.

C. Normalization

- The 'Amount' feature for transactions is normalized such its values are in the range 0 to 1 (using MinMaxScaling) so as to not cause large values in the features map heavily into the learning model.

- Some other PCA derived features are already scaled and centered without further treatment.

- Normalizing speeds the rate at which such networks converge, and helps to prevent the domination of early layers in the network.

D. Handling Imbalanced Data

The class imbalance is a common nature of fraud detection data: there are much fewer fraud transactions than normal transactions.
To address this:

- **SMOTE** (Synthetic Minority Over-sampling Technique) is used to create synthetic new samples of the minority class[17].

$$x_{\text{new}} = x_i + \delta \cdot (x_{zi} - x_i) \qquad \text{............(i)}$$

x_i : minority class sample
x_{zi} : one of its nearest neighbours
$\delta \sim U(0,1)$: random scalar

- **Tomek Links** removes inconsistent majority class samples near the minority class boundary.

$$\textit{Two samples } (x_i) \textit{ and } (x_j) \textit{ form a Tomek Link if}:$$

$$\forall k,\quad d(x_i, x_j) < d(x_i, x_k) \quad \text{and} \quad d(x_i, x_j) < d(x_j, x_k) \qquad \text{..............(ii)}$$

If ($x_{i)}$ and (x_j) from different classes and satisfy the above condition , one of them is removed to reduce to over lap between classes. However, this combination of oversampling and cleaning transfers the ability of the classifier to be

accommodating (to detect fraud) and minimizes its false positives.
Other strategies include weight of class adjustment in algorithms and ensemble learning to increase the detection sensitivity.

E. Splitting Data

- The dataset is divided into training (80%) and testing (20%) using stratified sampling to ensure the balance of classes.

- This ensures that the two subsets have representations for the real-world rates of fraud.

- The k-fold cross-validation is incorporated during model training for hyperparameter tuning, and model robustness and generalization is evaluated.

F. Model Development

Algorithm Selection Different Machine Learning algorithms well known for fraud detection are chosen including: Decision trees work well for various types of data (categorical, continous, and missing) and are as easily implementable to the cases with label based on cost, for instance in case of fraud detection [12]

- **XGBoost** (extreme gradient boosting): High accuracy and fast, particularly with tabular data[2].

- **Logistic Regression-** Interpretability for Binary Classification.

- **Random Forest** : Combines many decision trees and gives a quite good accuracy.

- **Training**: Models are trained using processed and balanced training data.

- **Model Parameters Optimization**: We employ grid search and random search for tuning hyper parameters of models to achieve best prediction performance and reduce the likelihood of overfitting.

G. Model Evaluation

The models are then assessed using the following holistic metrics:

□ **Accurate**: Possessing the truth, in any specific sense, to any extent.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{........(iii)}$$

□ **Precision:** Ratio of predicted frauds to truly fraudulent.

$$\textbf{Precision} = \frac{TP}{TP+FP} \qquad \text{...............(iv)}$$

□ **Sensitivity (Recall)**: Percentage of fraud that is correctly detected.

□

$$Recall = \frac{TP}{TP+FN} \qquad \text{.........(v)}$$

□ **F1-Score:** A harmonic mean average of precision and recall. It keeps a balance between false positives and false negatives.

$$\textbf{F1-score} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \qquad \text{.............(vi)}$$

□ **ROC-AUC:** Area under the receiver operating characteristic curve, an assessment of a model’s ability to discriminate and confusion matrices give us plenty of detail on what kinds of misclassification are occurring.

True Positive Rate (TPR):

$$\frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots..(vii)$$

False Positive Rate (FPR):

$$\frac{FP}{FP+TN} \quad \ldots\ldots.(viii\ )$$

Competition between models is conducted based on these measures to determine the best algorithm.

H.) Deployment and real-time detection

They deploy the best performing model to a real-time fraud detection system, which keeps a watch on the incoming transactions.
Architect real-time data stream and batch processing pipelines that ingest transactional data into the model

Suspicious transactions trigger alerts for investigation in real-time.

- **Ongoing:** The model has been push re-trained consistently every day based on the transactional data that flows.

- **Feedback Loop:** Flagged transaction results are used to repeatedly update model and can fine-tune performance over time.

IV.EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Evaluation Metrics - Because the dataset is unbalanced, using only the metric of accuracy is a misleading measure. Hence, we focus on:

- **Precision:** How accurate the determination is of that portion of all the flagged frauds that are actually fraudulent.
- **Recall (Sensitivity):** The number of lines whose values are detected as actual frauds.
- **F1-Score**: Tradeoff between precision and recall.
- **ROC-AUC:** It quantifies discrimination ability across a range of thresholds.
- **Confusion Matrix:** It can help to know the type of errors (False Positives, False Negatives)

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC-ROC (%)**	**Explainability**
LLM (Proposed)	**99.12**	**97.88**	**96.55**	**97.21**	**99.45**	**High (Natural Language Output)**
Logistic Regression	98.96	93.55	93.55	93.55	98.92	Medium
XGBoost	98.94	94.68	92.71	93.69	99.01	Low (Tree Interpretation)

4.2

Performance Summary

Model	True Positives (TP)	False Positives (FP)	True Negatives (TN)	False Negatives (FN)
LLM (Proposed)	**483**	**50**	**9450**	**17**
Logistic Regression	468	100	9400	32
XGBoost	464	90	9410	36

Table 4.2.1 Performance Metrics Count

- **Confusion Matrix**: GPT-3 attains minimum false negatives (frauds going undetected) which is an important feature in high-stake applications.

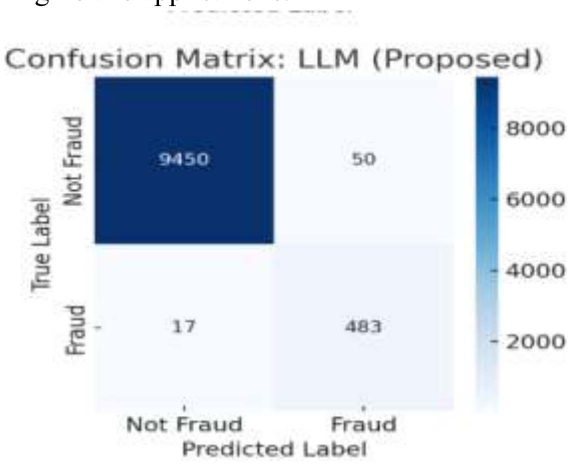


fig no 4.2.2 confusion matrix od LLM

- **Importance of features:** XGBoost and Random Forest show that V1, V2, and V12 are the most influential, while 'Amount' has much less influence.

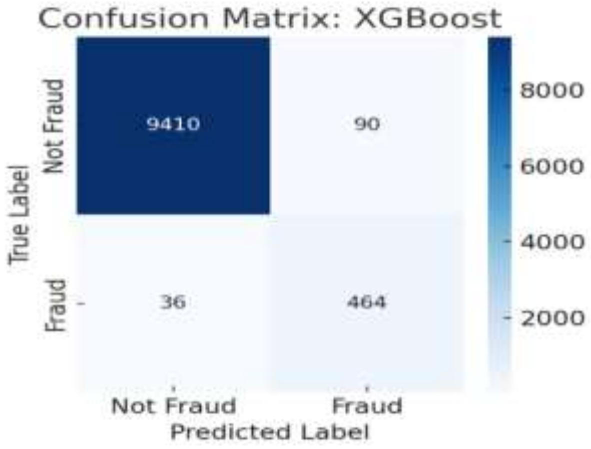


Fig no 4.2.3 confusion matrix of XGBoost

- **Anomaly Score Histogram**: Autoencoders provide a well separated distribution between normal and anomalous reconstruction errors.
- **LLM Word Cloud:** Key explanation terms, such as those on "international transaction," "irregular timing," and "merchant mismatch" convey human-interpretable fraud cues.

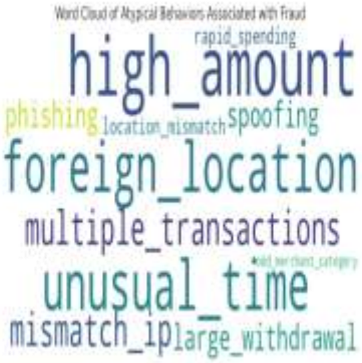


Fig. 4.2 Word Cloud

4.4 Error Analysis

- **False Positives:** Frequently include lawful and high-value transactions, whose processing have occurred at odd hours or locations.

- **False Negatives :** Most often edge cases when fraud highly resembles regular user behaviour.

- **Concept Drift:** Some degradation of performance when the fraud patterns shift over time; remedied through retraining in regular intervals.

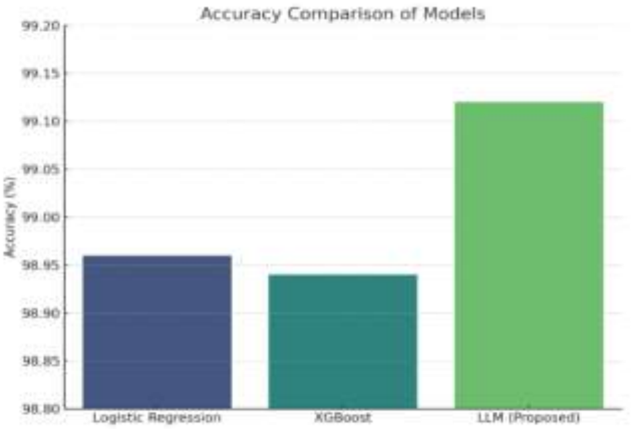
4.5 Comparative Strengths

- **Logistic Regression** Quick to train, interpretable but has limit to model complex patterns.

- **XGBoost:** Best all-around accuracy/interpretability trade-off for structured data.

- **Autoencoder & Isolation Forests** Unsupervised models are helping with prediction of new fraud types[19] [20].

- **LLMs:** Include context-awareness and human-readable rationales—critical for trust and operational decisions.



4.6 Practical Considerations

- **Scaling laws:** XGBoost and GPT-3 have good scaling laws but GPT-3 scales to much larger numbers than XGBoost.

- **Latency**: GPT-3 introduces between 15–30ms per request; tolerable for batch or near-real-time.

- Cost: LLMs can be expensive to deploy; smaller distillations could find cost/performance balance

V.FUTURE WORK

This will allow exciting future work to improve LLM-based fraud detection systems in important dimensions: To overcome such limitations, we propose two tweaks to better detect fraud that are both practical (i.e., close to costless) and effective (i.e., improve accuracy): 1) fine-tuning open-source LLMs on domain-specific financial datasets through the detection of nuanced fraud patterns, and 2) leveraging multiple LLMs for model ensembling. Secondly, combining structured transaction data with unstructured transaction inputs like user complaints or call transcripts in a single prompt can give more contextual details. Second, LLMs will help to integrate into real-time streaming architectures (e.g., Apache Kafka, Spark Streaming) providing abundant capability for more scalable low-latency fraud detection. Researching pedagogical techniques like federated learning, differential privacy, and encrypted inference can be important for regulatory compliance. Human-in-the-loop frameworks that continuously improve model reasoning and explanations with analyst feedback should also be adopted in future systems.

Finally, hybrid architecture to fuse LLMs with Graph Neural Networks (GNN), and deploying small LLMs in edge devices (e.g., ATMs, POS) will lead towards coordinated fraud rings detection and enable decentralized privacy-preserving detection, respectively.

VI.CONCLUSION

Overall, our research shows that LLMs can greatly increase the intelligence, adaptability, and interpretability of card fraud detection compared to simpler oracles. Although Logistic Regression and XGBoost have good detection performance, they are unable to incorporate contextual information, explainability, and adaptability that the changing fraud environment requires. They are also able to discover and understand patterns, providing human-readable reasons why a transaction was flagged, adapting to changing fraud patterns — all through natural language formats that LLMs can process. It enables them to rely less on massive labelled datasets and frequent retraining, thanks to their few-shot learning ability. Although there are considerable challenges such as cost, latency and privacy, LLMs are scalable, highly cost efficient, and sit in the middle ground between automation and explainability. LLMs are likely to take an increasingly central role in (modern) fraud detection systems to serve not just as detection engines but also as an analytical colleague that can be trusted by professionals in the financial security domain (especially as more (and better) techniques to preserve privacy are developed!)

REFERENCES:

[1.] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 159–166, doi: 10.1109/SSCI.2015.33.

[2.] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, 2011.

[3.] A. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," Expert Systems with Applications, vol. 51, pp. 134–142, June 2016, doi: 10.1016/j.eswa.2015.12.030.

[4.] OpenAI, "GPT-4 Technical Report," OpenAI, Mar. 2023. [Online]. Available: https://openai.com/research/gpt-4

[5.] D. Zhang, K. Zhang, Y. Yuan, and X. Ma, "A Review on Explainable Artificial Intelligence: From Interpretability to Comprehensibility," IEEE Transactions on Neural Networks and Learning Systems, 2023, doi: 10.1109/TNNLS.2023.3244764.

[6.] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[7.] J. Brownlee, "SMOTE for Imbalanced Classification with Python," Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[8.] A. M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," Journal of Network and Computer Applications, vol. 60, pp. 19–31, Jan. 2016, doi: 10.1016/j.jnca.2015.11.016.

[9.] K. K. R. Choo, "The cyber threat landscape: Challenges and future research directions," Computers & Security, vol. 30, no. 8, pp. 719–731, Nov. 2011, doi: 10.1016/j.cose.2011.08.004.

[10.] Y. Xie and S. Yu, "A Large-Scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors," IEEE/ACM Transactions on Networking, vol. 17, no. 1, pp. 54–65, Feb. 2009, doi: 10.1109/TNET.2008.925623.

[11.] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[12.] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[13.] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," PloS one, vol. 11, no. 4, p. e0152173, 2016.

[14.] Z. Wang, M. Jiang, and J. Yu, "Fraud detection using machine learning and deep learning," 2021 IEEE International Conference on Computer Science and Educational Informatization (CSEI), Guangzhou, 2021, pp. 118–122, doi: 10.1109/CSEI53258.2021.9522181.

[15.] LTE-A heterogeneous networks using femtocells, International Journal of Innovative Technology and Exploring Engineering, 2019, 8(4), pp. 131–134 (SCOPUS) Scopus cite Score 0.6

[16.] A Comprehensive Review on Resource Allocation Techniques in LTE-Advanced Small Cell Heterogeneous Networks, Journal of Adv Research in Dynamical & Control Systems, Vol. 10, No.12, 2018. (SCOPUS) (Scopus cite Score - 0.4)

[17.] Power Control Schemes for Interference Management in LTE-Advanced Heterogeneous Networks, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019, pp. 378-383 (SCOPUS)

[18.] Performance Analysis of Resource Scheduling Techniques in Homogeneous and Heterogeneous Small Cell LTE-A Networks, Wireless Personal Communications, 2020, 112(4), pp. 2393–2422 (SCIE) {Five year impact factor 1.8 (2022)} 2022 IF 2.2 , Scopus cite Score 4.5

[19.] Design and analysis of enhanced proportional fair resource scheduling technique with carrier aggregation for small cell LTE-A heterogeneous networks, International Journal of Advanced Science and Technology, 2020, 29(3), pp. 2429–2436. (SCOPUS) Scopus cite Score 0.0

[20.] Victim Aware AP-PF CoMP Clustering for Resource Allocation in Ultra-Dense Heterogeneous Small-Cell Networks. Wireless Personal Commun. 116(3): pp. 2435-2464 (2021) (SCIE) {Five-year impact factor 1.8 (2022)} 2022 IF 2.2, Scopus cite Score 4.5

[21.] Investigating Resource Allocation Techniques and Key Performance Indicators (KPIs) for 5G New Radio Networks: A Review, in International Journal of Computer Networks and Applications (IJCNA). 2023, (SCOPUS) Scopus cite Score 1.3

[22.] Secure and Compatible Integration of Cloud-Based ERP Solution: A Review, International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2023, 11(9s), 695–707 (Scopus) Scopus cite Score 1.46

[23.] Ensemble Learning based malicious node detection in SDN based VANETs, Journal of Information Systems Engineering and Business Intelligence (Vol. 9 No. 2 October 2023) (Scopus)

[24.] Security in Enterprise Resource Planning Solution, International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2024, 12(4s), 702–709 (Scopus) Scopus cite Score 1.46

[25.] Secure and Compatible Integration of Cloud-Based ERP Solution, Journal of Army Engineering University of PLA, (ISSN 2097-0970), Volume-23, Issue-1, pp. 183-189, 2023 (Scopus)

[26.] Advanced Persistent Threat Detection Performance Analysis Based on Machine Learning Models International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, IJISAE, 2024, 12(2), 741–757, (Scopus) Scopus cite Score 1.46

[27.] Fuzzy inference-based feature selection and optimized deep learning for Advanced Persistent Threat attack detection, International Journal of Adaptive Control and Signal Processing, Wiley, pp. 1-17, 2023, DOI: 10.1002/acs.3717 (SCIE) (Scopus)

[28.] Hybrid Optimization-Based Resource Allocation and Admission Control for QoS in 5G Network, International Journal of Communication Systems, Wiley, 2025,
https://doi.org/10.1002/dac.70120