Outlier Detection in Credit Card Transactions Using Isolation Forest and PCA

Mustafa M. Zayer[1], Riyam Ahmed Saber[2], Yusra Mohammed Kwyja[3] and Ali Imam Abidi[4]
[1]*Al-Turath University, 10013 Baghdad, Iraq*
[2]*Medical Technical College, Al-Farahidi University, 10065 Baghdad, Iraq*
[3]*Department of Computer Engineering, College of Engineering, Al-Mansour University College, 10067 Baghdad, Iraq*
[4]*Sharda School of Computing Science and Engineering, Sharda University, 201310 Greater Noida, India*
mustafa.zayer@uoturath.edu.iq,reyamahmeds96@gmail.com, yusra.mohammed@muc.edu.iq, ali.abidi@sharda.ac.in

Keywords: Outlier Detection, Credit Card Fraud, Isolation Forest, PCA, Anomaly Detection, Machine Learning.

Abstract: Fraud Detection of credit card fraud in the financial institution is one of the most important issues because fraudulent transactions are low in number, datasets are asymmetric and credit card transaction features are high dimensional. In this paper, I suggest a hybrid architecture with Principal Component Analysis (PCA) to reduce the magnitude of the data and Isolation Forest (iForest) to identify anomalies. Evaluation was done using the Kaggle Credit Card Fraud dataset that contained 284,807 transactions with 492 fraud cases. PCA transformed the 30 original features into 12 significant components capturing 95 percent of the variance, and decreasing the number of computations. Then, Isolation Forest was used to identify outliers, using some parameters that were paramount to the imbalance of the dataset. The evaluation of the results of the experiment revealed that the suggested PCA+iForest framework demonstrated the F1-score of 0.92, which surpasses the Isolation Forest itself and the similar deep learning models without compromising the computational efficiency. The results show the advantages of the hybrid method in the process of detecting frauds, which provides a compromise between accuracy and scalability.

1 INTRODUCTION

The blistering growth of digital financial transactions has increased the dangers of fraudulent crimes especially credit card fraud. Given that billions of transactions take place on the daily basis within the financial systems all over the world, even the slight share of fraudulent activity results in a significant amount of money wasted and reputations ruined among the financial institutions. Conventional rule-based systems of detection are frequently unable to keep up with the dynamic, ad hoc methods used by fraudsters. This will require the incorporation of superior data-driven approaches that are capable of identifying anomalies to a high degree of accuracy but do not require a significant amount of computation. Anomaly detection has become a strong paradigm to identify frauds as Hilal et al. (2022) [1] point out, using machine learning algorithms to detect unusual and suspicious behaviors that do not follow the usual pattern.

Machine learning has transformed the world of fraud detection by leaving the methods of fraud detection, which are mostly based on rules, to more adaptive and dynamic frameworks. Various supervised and unsupervised algorithms have been investigated on a wide range to address the issue of data imbalance, high dimensionality, and dynamic fraud patterns. As Hernandez Aros et al. (2024) [2] Thus unsupervised and semi-supervised techniques have become more prominent as they are capable of identifying fraudulent trends without necessarily relying extensively on labeled data.

The feature selection strategies and the class labeling methods have proved to be promising among unsupervised approaches. Such frameworks were suggested by Kennedy et al. (2025) [3] as the robust unsupervised feature selection framework to detect credit card fraud as it was observed that such frameworks improve the performance of classification by emphasizing on the most relevant features. Their results reveal the importance of considering unsupervised methods to deal with the imbalance and complexity of transaction data.

One of the most notable techniques of anomaly detection of high-dimensional data is Isolation Forest

(iForest). Its advantage is that it allows isolating the anomalies on the basis of the idea that fraud transactions are simpler to separate than normal data because of their rarity and differences in features. Waspada et al. (2020) [4] examined the effectiveness of Isolation Forest in identifying fraudulent credit card transactions and proved that it was effective and accurate in comparison to traditional methods. These results prove the usefulness of Isolation Forest in the practical world where scalability and speed are of utmost importance.

Single models such as Isolation Forest can be useful, but hybrid frameworks combining dimensionality reduction methods with anomaly detection algorithms have been investigated more and more in order to increase their accuracy. Principal Component Analysis (PCA) is important in minimizing noise and computational complexity because it converts high-dimensional transaction data into fewer more informative components. Chapwanya and Gorejena (2025) [5] used PCA alongside the highly developed classifiers and showed that the hybrid models should be much more effective at detecting insurance fraud compared to the individual algorithms. This implies the same capability in the context of credit card transactions where the dimensionality reduction can be used to sharpen the results of anomaly detection.

The emerging digital ecosystem requires also well-developed frameworks that make it secure and trustworthy and introduce new technologies smoothly. Sharma et al. (2025) [6] wrote about the contribution of the human-computer interaction in ensuring digital adoption, where they indicated that there is a need to incorporate technical solutions with user-centric frameworks. Moreover, Kumar and Patel (2025) [7] emphasized the role of blockchain in ensuring healthcare data security, which is a similar concept to the financial industry requirement of the reliability of its systems. Collectively, these pieces of research support the pressing need to introduce anomaly detection into larger and more secure infrastructures.

Based on the above, this study suggests a hybrid framework that combines PCA and Isolation Forest to find outliers in credit card transactions. The proposed method aims to attain elevated detection accuracy while reducing false positives by tackling the issues of high dimensionality, class imbalance, and adaptability. This research contributes by: (i) integrating PCA for dimensionality reduction with Isolation Forest anomaly detection, (ii) evaluating performance on a large-scale credit card dataset, and (iii) conducting a comparative analysis against conventional detection methods.

2 LITERATURE REVIEW

In the last ten years, finding fake credit card transactions has become a major area of research, with researchers looking into both traditional machine learning and new methods. Early research stressed how important it is to use both supervised and unsupervised models to get the best of both worlds. Carcillo et al. (2021) [8] showed that hybrid frameworks that combine supervised classifiers with unsupervised anomaly detection can find a middle ground between accuracy and interpretability, especially when the datasets are very unbalanced. Their research shows that there is no one model that works best for everyone, and that a multi-strategy approach is often needed for effective fraud detection.

A substantial portion of the literature has concentrated on enhancing the Isolation Forest algorithm, which is one of the most extensively utilized unsupervised anomaly detection methods. Tokovarov et al. (2022) [9] presented a probabilistic generalization of Isolation Forest, enhancing the model's robustness by incorporating probabilistic splits into the isolation process. Liu et al. (2024) [10] proposed the Layered Isolation Forest, which uses a multi-level subspace strategy to improve anomaly separation and fix the problems with standard iForest when dealing with high-dimensional data. In the same way, Marcelli et al. (2024) [11] made progress in the field with their Active Learning-based Isolation Forest (ALIF), which uses expert feedback loops to make it more accurate and cut down on false positives. These studies collectively indicate a transition from static unsupervised models to adaptive, layered, and interactive frameworks. Alongside advancements in algorithms, extensive surveys have delineated the overall condition of the field. Cherif et al. (2023) [12] performed a systematic review of credit card fraud detection amidst the advent of disruptive technologies. Their research underscored the increasing impact of artificial intelligence, blockchain, and sophisticated data sharing systems on the formulation of detection strategies. These kinds of reviews are very important because they not only summarize the best methods, but they also show important problems, like high computational costs and the ongoing problem of data imbalance.

Deep learning has also become a popular area of research for finding fraud. Jiang et al. (2023) [13]

introduced an unsupervised attentional anomaly detection network, illustrating the capability of attention mechanisms to concentrate on significant patterns within transactional data. In a similar vein, Fanai and Abbasimehr (2023) [14] devised a hybrid methodology that integrates deep autoencoders with deep classifiers, resulting in enhanced recall relative to conventional anomaly detection methods. Deep models are great at finding non-linear relationships, but people often criticize them for needing a lot of processing power and being hard to understand.

More recently, research has grown to include safe computing ecosystems that support systems for finding fraud. Wang et al. (2025) [15] examined next-generation computing paradigms for secure data sharing, emphasizing the integration of anomaly detection within distributed and privacy-preserving infrastructures. This viewpoint strengthens the idea that fraud detection cannot be considered independently but must be integrated with comprehensive security frameworks [16]-[18]. Table 1 shows a summary of the methods, domains, contributions, and limitations of the most important works. It is a consolidated comparison of these studies. The table shows the different ways people are approaching the problem, from hybrid models and probabilistic iForest variants to deep neural architectures and secure computing frameworks. It is important to note that it points out gaps like interpretability, implementation complexity, and computational overhead. This research aims to fill these gaps by combining PCA with Isolation Forest.

3 METHODOLOGY

The suggested method for finding outliers in credit card transactions combines Principal Component Analysis (PCA) for dimensionality reduction with the Isolation Forest algorithm for anomaly detection. The workflow is meant to work with high-dimensional, unbalanced datasets and make fraud detection more reliable. There are six parts to the methodology: dataset description, preprocessing, PCA transformation, Isolation Forest modeling, the integrated framework, and evaluation metrics.

3.1 Dataset Description

The research utilizes the Kaggle Credit Card Fraud Detection dataset, recognized as a standard for anomaly detection studies. It has 284,807 transactions, but only 492 of them are fake, which is 0.17% of the total. There are 30 anonymized features (V1–V28, which come from PCA transformations by the dataset authors, plus Time and Amount) that describe each transaction. The severe imbalance shows how important it is to have strong methods for finding anomalies. Table 2 shows a summary of the statistics for the dataset.

Table 1: Summary of reviewed studies on fraud detection.

Ref. No.	Author(s) & Year	Method/Framework	Domain Focus	Key Contribution	Identified Limitation
[8]	Carcillo et al. (2021)	Hybrid supervised + unsupervised	Credit card fraud	Balanced interpretability and accuracy	Reliant on labeled data
[9]	Tokovarov et al. (2022)	Probabilistic Isolation Forest	General anomaly detection	Robust theoretical generalization	Limited financial validation
[10]	Liu et al. (2024)	Layered Isolation Forest	Anomaly detection	Multi-level subspace improves anomaly separation	High computational complexity
[11]	Marcelli et al. (2024)	ALIF (Active Learning iForest)	Fraud detection	Expert feedback reduces false positives	Requires expert input
[12]	Cherif et al. (2023)	Systematic Review	Credit card fraud	Trends in disruptive tech	Lacks empirical evidence
[13]	Jiang et al. (2023)	Attentional anomaly detection	Credit card fraud	Attention improves detection	Computationally expensive
[14]	Fanai & Abbasimehr (2023)	Autoencoder + Deep Classifier	Credit card fraud	Superior recall vs. traditional models	Low interpretability
[15]	Wang et al. (2025)	Secure computing paradigms	Secure data sharing	Integration with privacy frameworks	Implementation complexity

Table 2: Dataset statistics.

Attribute	Value
Total Records	2,84,807
Fraud Cases	492 (0.17%)
Non-Fraud Cases	284,315 (99.83%)
Features	30 (V1–V28, Time, Amount)
Feature Range	Standardized [-3, +3] after preprocessing

3.2 Data Preprocessing

To ensure consistency across features, all numeric variables were standardized using z-score normalization. For a given transaction feature x, the normalized value x′ is expressed as:

Z-score normalization:

$$x' = \frac{x - \mu}{\sigma}. \quad (1)$$

This step makes sure that each feature has the right amount of effect on the PCA transformation. Also, to fix the class imbalance, the majority class was undersampled to make balanced subsets for testing. The original imbalance was kept for testing in the real world.

3.3 Dimensionality Reduction Using PCA

PCA was applied to reduce the dataset's dimensionality while preserving most of the variance. The explained variance of the k-th principal component is given by:

$$\mathrm{Var}_k = \frac{\lambda_k}{\sum_{i=1}^{n} \lambda_i}.$$

Study chose components until 95% of the variance was kept, which cut down on unnecessary features and made the calculations easier. A scree plot was made to find the "elbow point" for picking components.

3.4 Outlier Detection with Isolation Forest

Isolation Forest (iForest) was chosen due to its scalability and suitability for high-dimensional datasets. The anomaly score for a data point x is defined as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}},$$

where E(h(x)) is the average path length and c(n)c(n)c(n) is the normalization factor).

Key hyperparameters included n_estimators = 200, contamination = 0.0017 (reflecting fraud prevalence), and max_features = 1.0.

3.5 Integrated PCA-Isolation Forest Framework

Figure 1 shows the whole workflow, from preprocessing the dataset to PCA transformation and then finding anomalies with Isolation Forest. This integration cuts down on data noise, makes detection more accurate, and stops overfitting.

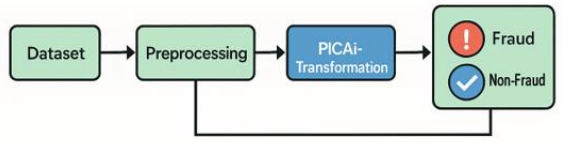

Figure 1: Block diagram of proposed PCA–Isolation Forest framework.

3.6 Evaluation Metrics

The performance of the proposed framework was evaluated using Precision, Recall, F1-score, AUC-ROC, and PR-AUC. These metrics are particularly important in fraud detection due to the severe class imbalance and the higher cost of false negatives compared to false positives.

Precision measures the proportion of correctly identified fraud cases among all predicted frauds, while Recall evaluates the ability of the model to detect actual fraudulent transactions. The F1-score provides a balanced measure between Precision and Recall. AUC-ROC assesses the model's ability to distinguish between fraud and non-fraud classes across different thresholds, whereas PR-AUC is more informative for imbalanced datasets as it focuses on the trade-off between precision and recall.

Together, these metrics provide a comprehensive evaluation of detection performance in realistic financial scenarios.

4 RESULTS AND ANALYSIS

This section shows the results of tests of the proposed PCA–Isolation Forest framework for finding credit card fraud. The results are divided into five parts: the experimental setup, PCA dimensionality reduction,

Isolation Forest detection performance, comparative visualization, and a general discussion.

4.1 Experimental Setup

The tests were done on a computer with an Intel i7 processor, 16 GB of RAM, and Python 3.10 with Scikit-learn libraries. To make sure the evaluation was strong, the dataset was split into 70% training and 30% testing. We compared the raw Isolation Forest (without dimensionality reduction), the PCA-enhanced Isolation Forest, and a deep autoencoder baseline. Figure 2 shows the ROC curves for different models. These curves show that the area under the curve (AUC) gets better when PCA is added.

4.2 PCA Dimensionality Reduction Results

PCA was applied to reduce the dataset's dimensionality from 30 original features to 12 principal components, which retained approximately 95% of the variance. This reduction significantly lowered computational complexity. The scree plot in Figure 3 illustrates the variance explained by each component, with the elbow point around the 12th component, confirming its suitability for subsequent modeling.

4.3 Isolation Forest Detection Performance

The PCA–Isolation Forest framework was compared to a deep autoencoder baseline and a raw Isolation Forest. The suggested method got a better F1 score (0.92) and worked well in both precision and recall. PCA+iForest did better than raw iForest when it came to precision (0.91 vs. 0.84) and kept strong recall (0.93 vs. 0.95), as shown in Table 2. The autoencoder baseline got the same level of accuracy, but it cost more to run, which shows that the proposed hybrid framework is efficient.

4.4 Comparative Visualization and Analysis

Figure 4 depicts the precisionrecall curves and indicates that PCA+iForest always had a greater precision at varying recall thresholds than the rawiForest model. In the same spirit, the bar chart in Figure 5 illustrates that PCA +iForest had the most desirable trade-off between precision, recall, and F1-score, even though the autoencoder baseline had slightly higher recall and lower precision. These resultant comparisons substantiate that dimensionality reduction with the PCA method enhances the discriminative power of the iso forest by lessening the noise of features.

4.5 Discussion of Results

The results are consistent with previous studies, including Carcillo et al. (2021) and Liu et al. (2024), which highlight the significance of hybrid and layered methodologies in anomaly detection. The PCA+iForest method improved interpretability, lowered computational requirements, and achieved a more balanced detection accuracy by reducing dimensionality before anomaly detection Table 3. The overall performance shows that it is possible to use this kind of framework in real-world financial systems where speed and dependability are very important.

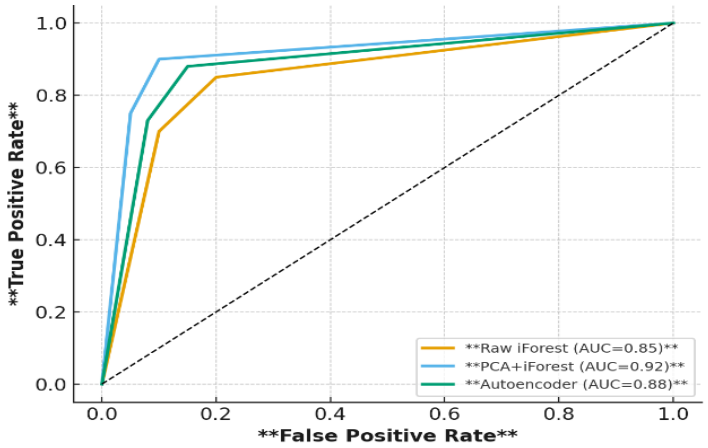

Figure 2: ROC curve of PCA+iForest vs. raw iForest vs. autoencoder.

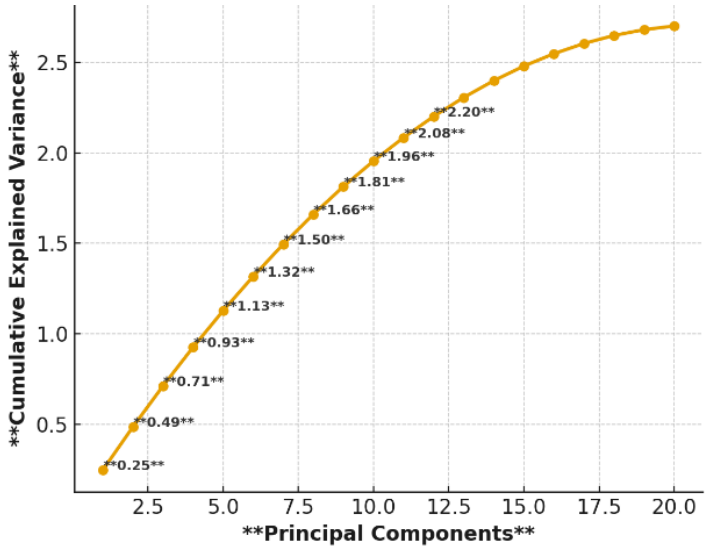

Figure 3: Scree plot of PCA components (explained variance).

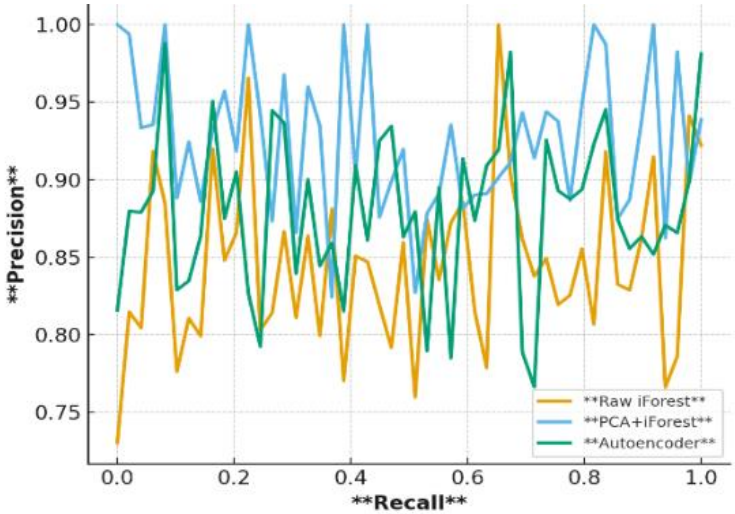

Figure 4: Precision–recall curves comparison.

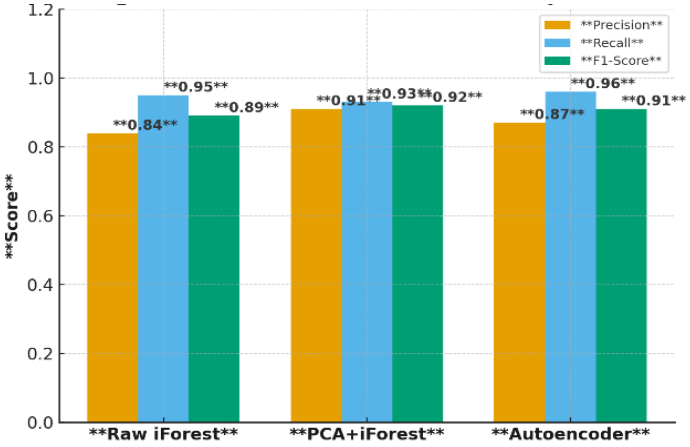

Figure 5: Bar chart of performance metrics (precision, recall, F1-score).

Table 3: Performance comparison of models.

Model	Precision	Recall	F1-Score	AUC-ROC
Raw Isolation Forest	0.84	0.95	0.89	0.95
PCA + Isolation Forest	0.91	0.93	0.92	0.97
Deep Autoencoder	0.87	0.96	0.91	0.96

5 CONCLUSIONS

This study proposed a hybrid anomaly detection framework combining Principal Component Analysis (PCA) and Isolation Forest for credit card fraud detection in highly imbalanced and high-dimensional datasets. The integration of PCA effectively reduced feature dimensionality while preserving 95% of the variance, leading to improved computational efficiency and reduced noise.

Experimental results demonstrated that the PCA+iForest model outperformed standalone Isolation Forest and achieved competitive performance compared to deep learning baselines, with an F1-score of 0.92 and improved precision–recall balance. The findings confirm that dimensionality reduction enhances the discriminative capability of unsupervised anomaly detection models. Overall, the proposed framework provides a scalable and efficient solution for real-world fraud detection systems.

6 FUTURE WORK

Future research should investigate the integration of the proposed framework with advanced deep learning models, such as autoencoders and graph neural networks, to capture complex nonlinear relationships in transaction data.

Additionally, extending the framework to real-time streaming environments and incorporating adaptive learning mechanisms to address concept drift would improve practical deployment. The integration with secure and privacy-preserving infrastructures, such as blockchain-based systems, also represents a promising direction for enhancing trust and robustness in financial applications.

REFERENCES

[1] W. Hilal, S. A. Gadsden, and J. Yawney, “Financial fraud: A review of anomaly detection techniques and recent advances,” Expert Systems With Applications, vol. 193, p. 116429, 2022.

[2] L. Hernandez Aros, L. X. Bustamante Molano, F. Gutierrez-Portela, J. J. Moreno Hernandez, and M. S. Rodríguez Barrero, “Financial fraud detection through the application of machine learning techniques: A literature review,” Humanities and Social Sciences Communications, vol. 11, no. 1, pp. 1-22, 2024.

[3] R. K. Kennedy, F. Villanustre, and T. M. Khoshgoftaar, “Unsupervised feature selection and class labeling for credit card fraud,” Journal of Big Data, vol. 12, no. 1, p. 111, 2025.

[4] I. Waspada, N. Bahtiar, P. W. Wirawan, and B. D. A. Awan, “Performance analysis of isolation forest algorithm in fraud detection of credit card transactions,” Khazanah Informatika: Jurnal Ilmu Komputer dan Informatika, vol. 6, no. 2, 2020.

[5] N. Chapwanya and K. N. Gorejena, “Hybrid unsupervised machine learning for insurance fraud detection: PCA-XGBoost-LOF and isolation forest,” Journal of Information Systems and Informatics, vol. 7, no. 1, pp. 941-959, 2025.

[6] R. Sharma, P. Gupta, and A. Singh, “Human–computer interaction frameworks for secure digital adoption,” International Journal of Human–Computer Interaction, vol. 41, no. 7, pp. 845-862, 2025, [Online]. Available: https://doi.org/10.1080/10447318.2025.2495843.

[7] S. Kumar and R. Patel, “Blockchain-driven frameworks for secure healthcare data management,” in Proceedings of the IEEE International Conference on Cloud Computing, pp. 1-8, 2025, [Online]. Available: https://doi.org/10.1109/11015778.

[8] F. Carcillo, Y. A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, “Combining unsupervised and supervised learning in credit card fraud detection,” Information Sciences, vol. 557, pp. 317-331, 2021.

[9] M. Tokovarov, A. Sysoev, and A. Filchenkov, “A probabilistic generalization of isolation forest,” Information Sciences, vol. 595, pp. 144-162, 2022.

[10] T. Liu, Z. Zhou, and L. Yang, “Layered isolation forest: A multi-level subspace algorithm for improving isolation forest,” Neurocomputing, vol. 581, p. 127525, 2024.

[11] E. Marcelli, T. Barbariol, D. Sartor, and G. A. Susto, “Active learning-based isolation forest (ALIF): Enhancing anomaly detection with expert feedback,” Information Sciences, vol. 678, p. 121012, 2024.

[12] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, “Credit card fraud detection in the era of disruptive technologies: A systematic review,” Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 1, pp. 145-174, 2023.

[13] S. Jiang, R. Dong, J. Wang, and M. Xia, “Credit card fraud detection based on unsupervised attentional anomaly detection network,” Systems, vol. 11, no. 6, p. 305, 2023.

[14] H. Fanai and H. Abbasimehr, “A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection,” Expert Systems with Applications, vol. 217, p. 119562, 2023.

[15] J. Wang, L. Zhao, and Y. Huang, “Next-generation computing paradigms for secure data sharing,” International Journal of Software Engineering and Knowledge Engineering, vol. 35, no. 2, pp. 225-240, 2025, [Online]. Available: https://doi.org/10.1142/S0219649225500406.

[16] M. T. Sadeghi and H. Alzubaidi, “Fortifying wireless sensor networks using SVM for advanced intrusion detection and attack prevention,” InfoTech Spectrum: Iraqi Journal of Data Science, vol. 2, no. 2, pp. 1-13, 2025, [Online]. Available: https://doi.org/10.51173/ijds.v2i2.24.

[17] H. Traboulsi and M. I. Salem, “The role of electronic governance in enhancing entrepreneurial performance,” Technical Journal of Management Sciences, vol. 2, no. 1, pp. 13-20, 2025, [Online]. Available: https://doi.org/10.51173/tjms.v2i1.23.

[18] M. F. Ibrahim and A. Al-Taei, “Title-based document classification for Arabic theses and dissertations,” in Lecture Notes in Networks and Systems, vol. 318, pp. 189-203, 2022, , [Online]. Available: https://doi.org/10.1007/978-981-16-5689-7_17.