# A MACHINE LEARNING-BASED FRAMEWORK FOR DETECTING FINANCIAL FRAUDULENT TRANSACTIONS

**Matheus H. R. Passos Hoffmann[1], Rhamon Alves Penteado[1], Marcos Monteiro Junior[1], Gabrielly de Queiroz Pereira[1]**

[1]Departamento de Informática – Universidade Estadual de Ponta Grossa (UEPG)
20011423@uepg.br, 21014223@uepg.br, gqpereira@uepg.br, mmjunior@uepg.br

***Abstract.*** *The objective of this study was to develop and evaluate machine learning models for the detection of financial fraud in transactions, aiming to mitigate economic losses and support decision-making in banking institutions. The methodology involved the use of the public Credit Card Fraud Detection dataset, consisting of 284,807 transactions, of which only 492 were fraudulent (≈ 0.172%). The algorithms Random Forest and XGBoost were tested, with and without the application of the SMOTE balancing technique. The evaluation was conducted using metrics such as precision, recall, F1-score, MCC, in addition to ROC and Precision–Recall curves. Complementarily, a qualitative validation was carried out through interviews with four financial sector specialists, in order to analyze the practical applicability of the models. The results showed that all models presented high overall performance, with areas under the ROC curve above 0.96. XGBoost with SMOTE achieved greater sensitivity, with a recall of 85% and 15 false negatives, but with an increase in false positives (22). On the other hand, Random Forest without SMOTE obtained better precision (0.94) and the highest F1-score (0.87), but failed to detect 18 fraud cases. Random Forest with SMOTE showed intermediate performance. The qualitative validation confirmed the relevance of the models, with 75% of the specialists prioritizing maximum fraud detection, even with more false alarms, and 25% valuing the reduction of false alarms, even at the cost of lower sensitivity. It is concluded that the choice of the model should consider the balance between recall and precision, aligned with institutional priorities between reducing financial losses and minimizing operational overload. The study also highlights limitations, such as the use of a specific temporal dataset and the absence of advanced hyperparameter optimization. For future work, it is suggested to explore parameter tuning, incremental learning, and validation on contemporary datasets, aiming for greater robustness and practical applicability of the models.*

## 1. INTRODUCTION

The exponential rise in digital financial transactions, coupled with the fast-paced evolution of fraudulent schemes, presents increasingly sophisticated challenges for banks, regulatory bodies, and end users. Although substantial investments have been made in security infrastructure, the vast amount and diversity of data produced render traditional rule-based detection systems inadequate, as they demand constant manual updates and fail to capture the complex, dynamic, and non-linear nature of fraudulent behavior [Sharma et al. 2021].

In this scenario, Machine Learning (ML) and Deep Learning (DL) techniques have emerged as powerful alternatives for automating fraud detection, capable of uncovering hidden patterns within large, heterogeneous datasets and identifying anomalies in near real time [Roseline et al. 2022]. Nevertheless, one of the main difficulties lies in the extreme imbalance of financial datasets, where legitimate transactions vastly outnumber

fraudulent ones. To address this issue, supervised resampling methods such as SMOTE and ADASYN have demonstrated substantial improvements in model performance by enhancing the representation of minority classes [Ileberi et al. 2021; Dang et al. 2021].

Moreover, research has shown that ensemble learning methods — including Random Forest, Gradient Boosting, and weighted voting ensembles — tend to outperform single models by integrating complementary classifier perspectives and reducing overall variance [Chhabra et al. 2023]. Hybrid architectures that incorporate attention mechanisms within LSTM or CNN frameworks have also achieved accuracy rates surpassing 99%, reinforcing the effectiveness of deep learning when combined with rigorous preprocessing procedures [Roseline et al. 2022; Jovanovic et al. 2022].

Building on these advances, this study aims to develop a predictive model for detecting financial fraud in credit card transactions, emphasizing three key components: (i) robust data preprocessing, including normalization, feature engineering, and class balancing; (ii) comparative assessment between traditional algorithms (Random Forest, SVM) and state-of-the-art models (XGBoost, deep neural networks); and (iii) careful metric selection — particularly recall, precision, and F1-score — to account for the operational costs of false negatives in real-world applications [Khan et al. 2022].

By integrating these stages into a replicable experimental framework, this research seeks to contribute to more reliable and adaptive fraud detection systems, promoting both financial loss reduction and greater confidence in digital banking environments.

## 2. METHODOLOGY

This section presents the procedures followed for the development, evaluation, and interpretation of predictive models aimed at detecting financial fraud. All stages were implemented in Python 3.11, using the libraries pandas, scikit-learn, imblearn, xgboost, matplotlib, seaborn, and shap to ensure efficiency and reproducibility throughout the experiments.

The study employed the Credit Card Fraud Detection public dataset, made available by the Université Libre de Bruxelles (ULB) on the Kaggle platform. The dataset contains 284,807 transactions carried out by European credit card holders in September 2013, of which 492 (≈0.172%) are labeled as fraudulent. The attributes include the anonymized PCA components V1–V28, the transaction amount in euros (Amount), the elapsed time in seconds since the first recorded transaction (Time), and the binary label (Class, where 0 = legitimate and 1 = fraudulent). The dataset was imported using the pandas.read_csv() function, maintaining all original records.

An exploratory data analysis (EDA) was conducted to understand data patterns and detect anomalies. Bar charts were used to visualize the imbalance between classes, histograms illustrated the distribution of transaction amounts, and temporal plots revealed the variation of the Time variable between legitimate and fraudulent operations. Outliers were examined through boxplots, while a correlation heatmap was generated to assess relationships among features.

For data preparation, the dataset was divided into 80% training and 20% testing subsets using a stratified approach (train_test_split, random_state=42), preserving the original fraud ratio in both partitions. The class imbalance was addressed using the SMOTE (Synthetic Minority Over-sampling Technique) method, applied exclusively to the training set to generate synthetic samples of the minority class. Since the variables V1–V28 were already standardized by PCA, no further normalization was required. The target variable Class was cast to the int8 type to optimize memory usage.

Three supervised classification models were trained for comparison:

- A baseline Random Forest without balancing (100 trees, default parameters).
- A Random Forest with SMOTE, identical to the baseline but trained on a balanced dataset;
- An XGBoost classifier with gradient boosting and eval_metric='logloss'.
- All models were trained with random_state=42 to guarantee reproducibility.

The evaluation of models was based on multiple performance metrics computed over the unbalanced test set, including the classification report (precision, recall, and F1-score for both classes), overall accuracy, and Matthews correlation coefficient (MCC). Additionally, the confusion matrix was analyzed to inspect false positives and negatives, and the ROC curve (AUC) and Precision–Recall curve (Average Precision) were generated to assess discriminative ability under high-class-imbalance conditions.

Beyond numerical evaluation, model interpretability was explored through feature importance and explainability techniques. The balanced Random Forest model provided feature importance based on impurity reduction across trees, visualized in horizontal bar charts. The SHAP library was employed via the TreeExplainer method to measure the contribution of each variable to individual predictions, summarized in a global importance plot highlighting the 20 most influential features.

Finally, a qualitative validation was conducted with financial-sector experts to assess the model's practical applicability. Semi-structured interviews were held with four professionals from distinct strategic areas — an Operations Director, a Risk and Internal Controls Manager, a Business Intelligence (BI) Manager, and a Development Superintendent — ensuring a broad organizational perspective. The interview guide addressed topics such as the impact of fraud on the institution, model suitability for real-time monitoring, trade-offs between high recall and high precision, adoption feasibility, and perceived risks. Participants also rated the model's practical utility on a five-point scale.

The responses were compiled into a comparative matrix, enabling descriptive quantitative analysis through frequency and percentage calculations, as well as identifying convergences and divergences among expert opinions. This qualitative phase complemented the experimental results, providing practical insights into the challenges and potential of deploying predictive fraud detection models in real-world financial environments.

## 3. RESULTS

This section presents the analysis of the outcomes obtained from the trained models for financial fraud detection. The experiments encompassed both traditional and modern supervised learning algorithms, emphasizing the impact of the SMOTE balancing technique on improving the identification of the minority (fraudulent) class. The performance graphs corresponding to the experiments are stored in files f1–f6 and referenced throughout this section.

The results were organized to show the evolution from the original dataset imbalance to model performance and expert validation.


**Figure 1. Distribuition of Classes (Fraudulent vs. non-Fraudulent)**

The first analysis highlighted the sharp imbalance between legitimate and fraudulent transactions, with frauds representing less than 0.2% of the total. This asymmetry was maintained in the stratified data split (80% training and 20% testing).

Such imbalance justified the use of resampling techniques like SMOTE, aiming to increase model sensitivity and ensure that fraudulent patterns were effectively learned.

**Table 1.**

| Dataset | Legitimate (Class 0) | Fraud (Class 1) | Fraud (%) |
|---|---|---|---|
| **Training** | 227,451 | 394 | 0.17 |
| **Testing** | 56,864 | 98 | 0.17 |

The prevalence of majority-class examples naturally biases most supervised models, underscoring the importance of balancing strategies to prevent the neglect of rare fraudulent cases.

Random Forest Performance (Without SMOTE)

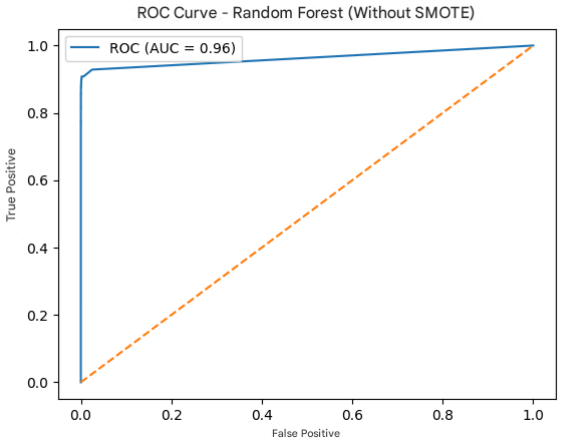


**Figure 2. ROC Curve – Random Forest (Without SMOTE)**

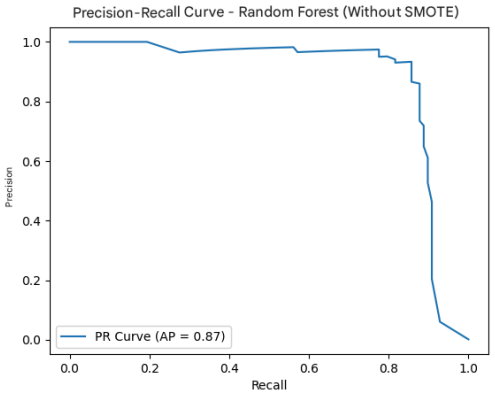


**Figure 3. Precision-Recall Curve – Random Forest (Without SMOTE)**

The baseline Random Forest trained on the original, unbalanced dataset achieved strong global performance, with AUC = 0.96 and Average Precision = 0.87. The Matthews Correlation Coefficient (MCC) reached 0.8763, indicating a robust relationship between predicted and actual labels.

However, the detailed analysis revealed 18 missed frauds (false negatives), which can be costly in operational environments. Although precision (0.94) and F1-score (0.87) were high for the minority class, the absence of SMOTE limited the model's ability to detect rarer fraudulent cases.

**Figure 4. ROC Curve – Random Forest (With SMOTE)**

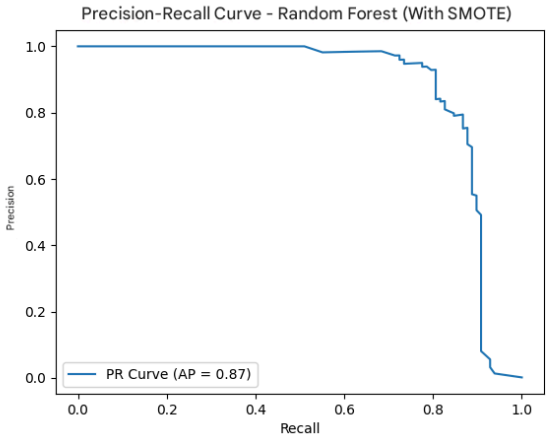


**Figure 5. ROC Curve – Random Forest (With SMOTE)**

After applying SMOTE, there was a slight improvement in recall for the fraudulent class (from 82% to 83%), reducing false negatives to 17. However, false positives rose from 5 to 17, lowering precision to 0.83.

The AUC remained at 0.96, while the MCC decreased slightly to 0.8262, indicating a minor trade-off between sensitivity and overall correlation. In essence, SMOTE enhanced the model's detection capacity but introduced more false alarms.

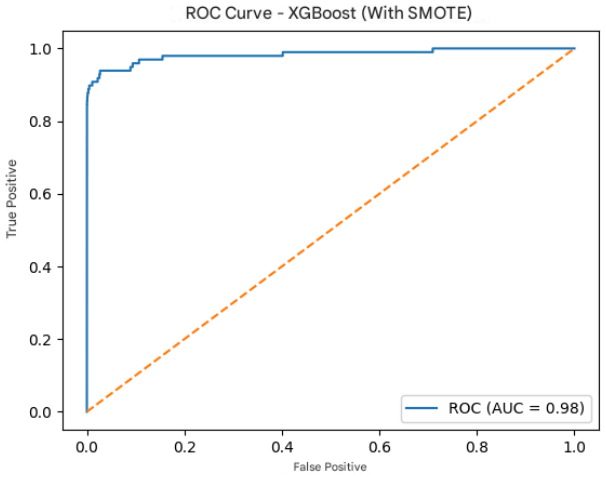


**Figure 6. ROC Curve – XGBoost (With SMOTE)**

Among the evaluated models, XGBoost with SMOTE achieved the best recall for fraudulent transactions (85%), missing only 15 fraud cases. However, the number of false positives increased to 22, resulting in a precision of 0.79.

The model's AUC reached 0.98, the highest across all tests, while MCC = 0.8179. This suggests that XGBoost is especially suited for high-risk environments where missing a fraud is more critical than handling a few additional false alarms.

**Table 2.**

| Model | Precision | Recall | F1-score | MCC |
|---|---|---|---|---|
| **Random Forest (no SMOTE)** | 0.94 | 0.82 | 0.87 | 0.8763 |
| **Random Forest (with SMOTE)** | 0.83 | 0.83 | 0.83 | 0.8262 |
| **XGBoost (with SMOTE)** | 0.79 | 0.85 | 0.82 | 0.8179 |

The comparative synthesis highlights key insights:

- All models achieved overall accuracy above 99.9%, but this metric is unreliable in highly imbalanced datasets.
- The Random Forest without SMOTE provided the best equilibrium between precision and F1-score, though it failed to detect some fraudulent cases.
- The XGBoost with SMOTE offered the highest recall, identifying most fraudulent operations at the expense of a higher false-positive rate.
- The SMOTE technique proved effective in increasing model sensitivity but required careful calibration to balance the operational cost of false alarms.

In practice, the model selection should depend on the trade-off between sensitivity (recall) and precision, considering the operational cost of false negatives and false positives. In critical financial monitoring scenarios, prioritizing higher recall—even with manual verification overhead—justifies the adoption of balanced XGBoost.

Beyond numerical evaluation, a qualitative validation was conducted through interviews with four financial-sector specialists — an Operations Director, a Risk and Internal Controls Manager, a Business Intelligence Manager, and a Development Superintendent. The purpose was to assess the practical applicability of the fraud detection model.

The main findings were:

- Fraud impact: 75% rated it as high, and 25% as medium.
- Model adequacy: 75% found the model suitable for fraud detection, while 25% considered it partially adequate.
- Application priorities: 75% preferred maximizing fraud detection even with more false alarms; 25% favored minimizing false alarms, accepting some missed cases.
- Adoption feasibility: 75% answered yes, and 25% maybe.
- Main risks: customer dissatisfaction due to false positives, need for trained staff to handle alerts, reputation management, and the fast evolution of fraud techniques.

- Suggested improvements: enrich datasets to reduce false alarms, include client behavior context, enable easier reversal of false-positive blocks (e.g., biometrics), and implement continuous-learning AI.
- Perceived usefulness: 100% rated the model as 4 or 5 out of 5 (50% each).

The chart above summarizes the experts' responses, showing an overall positive perception toward adopting the model, despite concerns related to the operational cost of false positives. The qualitative feedback reinforced the experimental conclusions, indicating that the proposed model is both technically sound and practically viable for real-world financial monitoring environments.

## 4. DISCUSSION

The results achieved in this study demonstrate a high overall performance across all tested models, with ROC curve areas exceeding 0.96 in every configuration. This finding indicates that, in general, the algorithms were able to effectively discriminate between legitimate and fraudulent transactions. However, when focusing on the minority class — composed of fraudulent transactions — important distinctions emerge that warrant deeper analysis. In fraud detection contexts, global accuracy alone is insufficient, as the strong class imbalance can conceal the models' inability to recognize rare yet financially significant events.

The XGBoost model combined with the SMOTE balancing technique achieved the highest sensitivity, reaching a recall of 85% and missing only 15 fraudulent cases. This demonstrates that training on a rebalanced dataset enhances the algorithm's ability to detect the target class. Nevertheless, this improvement in recall was accompanied by a considerable rise in false positives (22 cases), indicating an operational trade-off. While a greater number of frauds are detected — reducing financial losses — institutions may face increased verification workload, higher operational costs, and potential customer dissatisfaction due to incorrect alerts.

By contrast, the Random Forest trained without SMOTE exhibited the best precision (0.94) and highest F1-score (0.87), achieving a more balanced compromise between precision and recall. However, it failed to identify 18 fraudulent transactions, reflecting a lower sensitivity to minority patterns. This model's conservative nature reduces false alarms but risks overlooking significant fraudulent behavior. The Random Forest with SMOTE showed intermediate behavior — slightly improving fraud detection (one less false negative) but producing more false positives (17), which decreased precision to 0.83. This comparison reinforces the classical precision–recall trade-off, where gains in sensitivity often entail a decline in precision.

An examination of the ROC and Precision–Recall curves helps clarify this dynamic. Although all models performed well globally, differences became more apparent in low false-positive regions. The XGBoost with SMOTE acted more aggressively in detecting frauds, while the unbalanced Random Forest maintained a conservative posture, avoiding unnecessary alerts but missing some anomalies. Therefore, the choice between models must align with institutional priorities: organizations focused on minimizing

financial risk should favor high-recall models, while those emphasizing customer experience and operational stability might prefer higher-precision models.

Feature importance analysis further revealed that variables V14, V17, and V12 were the most influential in decision-making. These findings are consistent with prior research identifying similar PCA-transformed variables as critical indicators of fraud. The identification of such features reinforces the reliability of the models and highlights the importance of explainable and interpretable AI, especially in the financial sector, where algorithmic transparency is increasingly required by regulatory frameworks.

The qualitative validation with financial experts complemented the quantitative analysis and provided insights into real-world applicability. Three of the four interviewees emphasized that the primary goal should be maximizing fraud detection, even at the cost of more false positives — an opinion consistent with the balanced XGBoost performance. The fourth expert, however, prioritized reducing false alarms, aligning with the Random Forest without SMOTE. This divergence of perspectives underscores that the optimal model depends heavily on the institution's operational goals and tolerance for error.

Experts also raised practical concerns regarding implementation, such as the reputational risk of incorrectly flagging customers, the need for specialized teams and infrastructure to manage alerts, and the importance of preserving user experience when resolving suspected frauds. Their feedback introduces an organizational dimension that purely quantitative metrics cannot capture. Suggested improvements included adopting continuous learning systems capable of adapting to evolving fraud tactics and customizing decisions according to client behavior profiles, considering factors such as income, region, and spending habits. These recommendations emphasize the necessity of aligning technological solutions with operational and user-centered goals.

A particularly encouraging outcome from the qualitative validation was the positive perception of the model's practical utility. All participants rated it 4 or 5 out of 5, indicating strong approval and perceived applicability despite the noted challenges. This endorsement strengthens the quantitative results, suggesting that the model could be effectively deployed in production environments — provided it is accompanied by organizational adjustments and ongoing monitoring.

Despite the promising findings, certain limitations must be acknowledged. The dataset represents a specific context — European credit card transactions from September 2013, anonymized through PCA — which restricts both interpretability and representativeness. Moreover, hyperparameter optimization techniques were not applied, potentially limiting performance gains. Finally, operational costs associated with false positives and false negatives were not incorporated into the evaluation metrics, even though they hold substantial real-world relevance.

Future work should explore hyperparameter tuning, incremental or real-time learning, and the inclusion of contextual behavioral features. Validation using more recent and representative datasets, ideally in collaboration with financial institutions, would further enhance reliability and applicability.

In summary, the findings indicate that the choice between Random Forest and XGBoost, with or without SMOTE, must balance sensitivity and precision while considering both operational costs and user experience. XGBoost with SMOTE emerged as the most effective for fraud detection, whereas Random Forest without SMOTE showed greater restraint in generating alerts. The qualitative validation confirmed this duality, demonstrating that model adoption must be tailored to each institution's risk tolerance and strategic objectives. By integrating quantitative and qualitative perspectives, this research provides a comprehensive understanding that combines technical performance with practical insights — paving the way for more effective applications of machine learning in financial fraud detection.

## 5. CONCLUSION

This study presented the development and evaluation of machine learning models for financial fraud detection, with particular emphasis on the effects of class balancing using the SMOTE technique. The results demonstrated that all models achieved ROC curve areas above 0.96, confirming their strong discriminative power. However, significant differences were observed regarding sensitivity and precision.

The XGBoost model with SMOTE achieved the highest sensitivity, reducing false negatives to 15 cases and reaching a recall of 85%, although this improvement came with an increased number of false positives. In contrast, the Random Forest without SMOTE maintained higher precision (0.94) and an F1-score of 0.87, yet failed to identify 18 fraudulent transactions. The Random Forest with SMOTE exhibited intermediate performance, confirming the classic trade-off between maximizing fraud detection and minimizing false alarms.

The qualitative validation supported these findings: most financial experts prioritized maximizing fraud detection, even at the cost of additional false positives, while a smaller group favored reducing incorrect alerts to minimize operational and reputational risks. The interviews also revealed practical concerns, such as the potential impact on institutional reputation, increased workload for fraud analysis teams, and the need for strategies that support continuous learning and client-specific customization.

Overall, the findings highlight that the choice of model must consider both technical metrics and organizational factors. Future work should focus on hyperparameter optimization, integration of contextual and behavioral variables, and validation with more recent and representative datasets, in order to enhance the model's robustness, adaptability, and real-world applicability.

## REFERENCES

Afjal, M., Salamzadeh, A., and Dana, L.-P. (2023). Financial fraud and credit risk: Illicit practices and their impact on banking stability. In: Journal of Risk and Financial Management, 16(9):386.

Alarfaj, F. K., Alotaibi, R., Almutairi, A., and Alotaibi, A. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. In: IEEE Access, 10:39700–39715.

Alfaiz, N. S. and Fati, S. M. (2022). Enhanced credit card fraud detection model using machine learning. In: Electronics, 11(4).

Alkurdi, A. A. H., Alenezi, M. N., and Albahar, M. A. (2024). Evaluating the impact of point-biserial correlation-based feature selection on machine learning classifiers: A credit card fraud detection case study. In: Revista Gestão & Tecnologia – Journal of Management and Technology, 24(2):166–196.

Chhabra, R., Goswami, S., and Ranjan, R. K. (2023). A voting ensemble machine learning-based credit card fraud detection using highly imbalanced data. In: Multimedia Tools and Applications.

Dang, T. K., Hoang, D. H., Nguyen, H. T., and Tran, M. T. (2021). Machine learning based on resampling approaches and deep reinforcement learning for credit card fraud detection systems. In: Applied Sciences – Basel, 11(21).

Dastidar, K. G., Caelen, O., and Granitzer, M. (2024). Machine learning methods for credit card fraud detection: A survey. In: IEEE Access, 12:158939–158965.

Feng, X. and Kim, S.-K. (2024). Novel machine learning-based credit card fraud detection systems. In: Mathematics, 12(12).

Gupta, S., Choudhury, T., Singh, R., and Alam, M. (2022). A hybrid machine learning approach for credit card fraud detection. In: International Journal of Information Technology Project Management, 13(3).

Ileberi, E., Sun, Y., and Wang, Z. (2021). Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost. In: IEEE Access, 9:165286–165294.

Jovanovic, D., Vukovic, S., and Milosevic, N. (2022). Tuning machine learning models using a group search firefly algorithm for credit card fraud detection. In: Mathematics, 10(13).

Khalid, A. R., Ahmad, I., and Malik, R. (2024). Enhancing credit card fraud detection: An ensemble machine learning approach. In: Big Data and Cognitive Computing, 8(1).

Khan, S., Rahman, M. M., and Alam, M. (2022). Developing a credit card fraud detection model using machine learning approaches. In: International Journal of Advanced Computer Science and Applications, 13(3):411–418.

Malik, E. F., Hassan, M., and Tariq, A. (2022). Credit card fraud detection using a new hybrid machine learning architecture. In: Mathematics, 10(9).

Mienye, I. D. and Sun, Y. (2023). A machine learning method with hybrid feature selection for improved credit card fraud detection. In: Applied Sciences – Basel, 13(12).

Ming, R., Zhang, T., and Liu, H. (2024). Enhancing fraud detection in auto insurance and credit card transactions: A novel approach integrating CNNs and machine learning algorithms. In: PeerJ Computer Science, 10.

Mosa, D. T., Al-Batati, M. A., and Saleh, M. (2024). CCFD: Efficient credit card fraud detection using meta-heuristic techniques and machine learning algorithms. In: Mathematics, 12(14).

Plakandaras, V., Gogas, P., Papadimitriou, T., and Tsamardinos, I. (2022). Credit card fraud detection with automated machine learning systems. In: Applied Artificial Intelligence, 36(1).

Roseline, J. F., Varghese, A., and Jose, A. (2022). Autonomous credit card fraud detection using machine learning approach. In: Computers & Electrical Engineering, 102.

Sharma, P., Kumar, A., and Gupta, S. (2021). Machine learning model for credit card fraud detection – a comparative analysis. In: International Arab Journal of Information Technology, 18(6):789–796.

Zhao, Z. and Bai, T. (2022). Financial fraud detection and prediction in listed companies using SMOTE and machine learning algorithms. In: Entropy, 24(8):1157.