Application of Machine Learning in Financial Fraud Detection and Prevention - A Comparative Analysis of Algorithms –

Adaleta Hasanović[1], Savo Stupar[2], Kemal Kačapor[3], Nijaz Bajgorić[4]*

[1]EY GmbH & Co. KG Wirtschaftsprüfungsgesellschaft, Germany, Düsseldorf
adaletahasanovic2701@gmail.com
[2]University of Sarajevo, School of Economics and Business, Bosnia and Herzegovina, Sarajevo
savo.stupar@efsa.unsa.ba
[3]University of Sarajevo, School of Economics and Business, Bosnia and Herzegovina, Sarajevo
kemal.kacapor@efsa.unsa.ba
[4]University of Sarajevo, School of Economics and Business, Bosnia and Herzegovina, Sarajevo
nijaz.bajgoric@efsa.unsa.ba*
*Corresponding author

Abstract: Detecting fraudulent activities in the financial sector is a critical challenge that requires robust, adaptive approaches. This paper investigates the application of machine learning (ML) algorithms—Logistic Regression, SVM, KNN, Decision Trees and Random Forests—for credit card fraud detection. Utilizing a highly imbalanced dataset, models were evaluated using precision, recall, and F2 score, prioritizing recall to minimize undetected fraud. Our findings demonstrate that Logistic Regression achieved the highest recall (91%), effectively identifying the majority of fraudulent transactions while maintaining a low false-negative rate. SVMs achieved balanced performance with 89% recall, while Random Forests showed superior precision (98%), minimizing false alarms. These results highlight the strengths and trade-offs of ML algorithms for uncovering complex patterns in large-scale financial data and for reducing fraud risk when integrated with real-time detection systems. This research underscores the importance of continuous model optimization using updated data and advanced techniques to counter evolving fraud tactics. By bridging technological innovation with

G. Szabó-Szentgróti et al. (eds.), *Proceedings of the Kautz Conference on Business and Economics 2025 (KCBE 2025)*, Advances in Economics, Business and Management Research 381,
https://doi.org/10.2991/978-94-6239-658-6_2

proactive fraud prevention, this paper provides actionable insights for financial institutions, contributing to the development of secure and resilient financial ecosystems.

Keywords: Machine learning, Credit card fraud, Fraud prevention, Logistic regression

1. Introduction

The rapid digitalization of financial services has transformed the way transactions are conducted, offering unparalleled convenience to consumers and businesses. However, this evolution has also created new vulnerabilities, enabling increasingly complex and covert forms of financial fraud. Among the most prevalent threats is credit card fraud, which poses significant risks to both consumers and institutions, particularly as online and card-not-present transactions continue to grow. Fraudsters are continually refining their techniques, making early and accurate detection of fraudulent behavior both essential and increasingly difficult.

Traditional fraud detection systems, often based on static, rule-based mechanisms, have proven inadequate in this evolving landscape. These systems tend to struggle with novel fraud patterns, resulting in either missed fraud cases (false negatives) or excessive false alarms (false positives), both of which lead to operational inefficiencies and reduced customer trust. In response, financial institutions are increasingly turning to machine learning (ML) methods, which provide more adaptive, data-driven solutions capable of detecting subtle and emerging patterns in vast volumes of transactional data.

Despite the advancements, a persistent challenge in fraud detection research remains: the severe imbalance between fraudulent and legitimate transactions. This imbalance can bias model learning and reduce effectiveness in detecting rare fraudulent activity. Furthermore, many existing studies focus either on single-model evaluations or do not account sufficiently for performance trade-offs between recall, precision, and the business costs of classification errors.

This paper aims to address these gaps through a comprehensive, comparative analysis of five widely adopted machine learning algorithms—Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Random Forests—in the context of real-world credit card fraud detection. Using a publicly available, highly imbalanced dataset, the paper evaluates each model based on performance metrics particularly suited to fraud detection scenarios, including recall,

precision, and the F2-score. The primary objective is to identify the algorithm that best balances the need to minimize undetected fraud (high recall) while maintaining acceptable false positive rates.

The contribution of this research lies in its practical focus on evaluating algorithmic trade-offs in a realistic setting, its emphasis on rare-event classification metrics, and its relevance to industry practitioners seeking to implement machine learning in operational fraud detection environments. Unlike many prior studies, this does not treat model selection as a purely technical task; instead, it aligns evaluation with the risk-sensitive nature of financial decision-making and discusses the implications of each model's strengths and limitations for real-world deployment. By systematically analyzing model behavior on an imbalanced dataset, this paper not only informs algorithm selection but also offers deeper insight into how machine learning can be effectively applied to enhance the resilience of financial systems against fraud.

The manuscript is organized as follows. Section contais a short literature review on machine learning in fraud detection and introduces the CRISP-DM methodology as the guiding framework. Sections 3 - 7 apply this framework to a practical case study on credit card fraud detection, covering the phases of Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. The results of the comparative analysis of five machine learning algorithms are presented and discussed. Finally, Section 8 concludes the paper by summarizing key findings, highlighting contributions, and outlining directions for future research.

2. Literature Review

Recent literature underscores the growing consensus on the effectiveness of machine learning in this domain. Ashtiani and Raahemi (2022) highlight the dominance of supervised learning approaches in financial fraud detection, particularly in structured datasets where historical labels enable pattern recognition. Aslam et al. (2022) demonstrate that ML and AI-based systems consistently outperform traditional methods, especially in minimizing false classifications—a critical concern in financial security. Additionally, Bin, Vitaly, and Paul (2022) emphasize the reliability of widely used algorithms such as logistic regression, decision trees, support vector machines, and random forests, particularly in the context of highly imbalanced datasets.

Several approaches in applying ML techniques in Credit Card Fraud Detection are presented in Ashtiani & Raahemi (2022), Bin, Vitaly & Paul (2022), Mahmood et al. (2023), Ileberi et al. (2021, 2022), Mutemi & Bacao (2024), Ramzan & Ahmed (2022).

Applying ML techniques in finance are discussed in Dowling, Aziz, & Hammami (2019), Ashtiani, M. N., & Raahemi, B. (2022) while Aslam et al. (2022) and Rukhsar et al. (2022) considered the application of AI and ML in insurance fraud detection. Guo et al. (2024) analyzed the application of machine learning in risk management.

Recent studies have demonstrated the effectiveness of the CRISP-DM framework in fraud detection applications. For instance, Pahuja and Kamal (2023) applied CRISP-DM to develop an ensemble learning model for detecting fraudulent transactions on the Ethereum blockchain, achieving high accuracy through systematic data preprocessing and model evaluation. Similarly, a paper published in Finance Research Letters (2023) utilized CRISP-DM to assess the risk profiles of fintech companies, facilitating the classification and clustering of algorithms and regression models in the evaluation process. By structuring this research around the CRISP-DM methodology, the ensures methodological transparency and alignment between technical outcomes and the practical needs of fraud detection in the financial sector.

We use the Cross-Industry Standard Process for Data Mining (CRISP-DM) as a framework for structuring the machine learning pipeline. CRISP-DM is a widely accepted methodology that provides a structured and iterative approach to data mining and analytics projects, aligning technical analysis with business objectives. Its applicability has been demonstrated in various domains, including fraud detection and risk assessment in financial sectors.

The CRISP-DM framework comprises six interrelated phases:

- Business Understanding: This initial phase focuses on defining the problem from a business perspective, identifying project goals, and determining success criteria. In fraud detection, this involves understanding the cost implications of false negatives versus false positives and prioritizing appropriate performance metrics.
- Data Understanding: This phase involves data collection, exploratory analysis, and assessment of data quality. Identifying class imbalance, feature distributions, and potential anomalies are critical steps in preparing fraud detection models.
- Data Preparation: This phase transforms raw data into a clean and structured form suitable for modeling. Tasks include handling missing values, normalizing features, balancing datasets (e.g., using SMOTE), and selecting relevant features. Effective data preparation directly influences model performance in imbalanced scenarios like financial fraud.
- Modeling: Machine learning algorithms are applied during this stage, with careful parameter tuning to optimize performance. Multiple models may be evaluated,

including logistic regression, decision trees, support vector machines, and ensemble methods.

- Evaluation: This phase assesses whether the models meet the business goals defined earlier. In fraud detection, this typically involves metrics such as recall, precision, F2-score, and confusion matrices to account for rare-event classification.
- Deployment: Although not executed in this paper, the final CRISP-DM phase involves deploying the model into a production environment, monitoring performance, and updating the model as fraud patterns evolve.

3. Business understanding

The business understanding phase is the initial step in the CRISP-DM methodology (Chapman et al., 2000). This phase focuses on defining project objectives and determining requirements from a business perspective.

From the exploratory perspective, the CRISP-DM process will conclude with model evaluation and selection, rather than real-world deployment in a financial institution. The phase includes defining the business objective and appropriate evaluation metrics for assessing model performance.

The dataset used in this paper was obtained from Kaggle (Machine Learning Group – ULB, 2019). It consists of 284,807 credit card transactions conducted by European cardholders in September 2013. The dataset includes 30 numerical features, such as time, transaction amount, and anonymized features (V1–V28) derived using Principal Component Analysis (PCA) to ensure data confidentiality. The target variable, "Class," labels transactions as either fraudulent or legitimate, enabling the application of supervised learning algorithms.

Fraudulent transactions account for 0.17% of all transactions (492 out of 284,807), indicating a highly imbalanced dataset. This class imbalance poses challenges for training machine learning models. Consequently, selecting appropriate evaluation metrics is crucial to emphasize rare observations (fraudulent transactions) and ensure a reliable assessment of model performance.

An imbalanced dataset occurs when one class is significantly underrepresented compared to another. In the context of credit card fraud detection, fraudulent transactions make up only 0.17% of the dataset (492 out of 284,807 transactions), while legitimate transactions account for 99.83% of the total. Research indicates that machine learning models often struggle to accurately detect the minority class in such datasets

(He & Garcia, 2009). This issue arises because the model perceives the minority class as statistical noise, reducing its predictive relevance. As a result, models trained on imbalanced data tend to classify most transactions as legitimate. When evaluated using an inappropriate metric, such as accuracy, these models may appear highly effective despite failing to detect fraudulent transactions.

Accuracy, defined as the ratio of correctly predicted transactions to the total number of transactions, becomes an unreliable metric in fraud detection. Given the high prevalence of legitimate transactions (99.83%), a model that classifies all transactions as non-fraudulent would achieve an accuracy of 99.83% without effectively identifying fraudulent transactions. Consequently, relying solely on accuracy can lead to misleading conclusions and the deployment of ineffective fraud detection systems, exposing financial institutions to financial losses and reputational risks (He & Garcia, 2009).

From a business perspective, the primary objective is to develop a machine learning model capable of reliably identifying fraudulent transactions while minimizing false positives and false negatives. However, the class imbalance complicates model evaluation, necessitating the use of appropriate performance metrics. To address these challenges, alternative evaluation metrics such as precision, recall, and F1-score should be used instead of accuracy (Khemakhem, 2020). These metrics provide a more comprehensive assessment of model performance by considering both the majority and minority classes and accounting for false positive and false negative errors, which are particularly critical in fraud detection:

- Precision measures the proportion of correctly identified fraudulent transactions among all transactions predicted as fraudulent.

- Recall measures the proportion of correctly identified fraudulent transactions among all actual fraudulent transactions.

- F1-score, the harmonic mean of precision and recall, provides a balanced measure of the model's effectiveness in detecting fraud.

To further analyze model performance, a confusion matrix will be introduced, which illustrates the relationship between predicted classifications and actual transaction labels (see Table 1).

Table 1 *Confusion Matrix with Four Possible Outcomes*

Actual Class	Predicted Fraud (+)	Predicted Non-Fraud (-)
Fraud (+)	True Positive (TP)	False Negative (FN)
Non-Fraud (-)	False Positive (FP)	True Negative (TN)

A more detailed explanation of the aforementioned metrics is provided below:

1. **Accuracy**: Accuracy refers to the proportion of correct predictions made by the model relative to the total number of predictions. It is a general measure of the model's overall performance. However, in imbalanced datasets, accuracy can be misleading as it does not differentiate between the different types of errors. Accuracy is defined as:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

2. **Precision:** Precision measures the proportion of true positive predictions among all instances that were predicted as positive by the model. In other words, it indicates how many of the predicted positive cases were actually correct. Precision is defined as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

3. **Recall:** Recall, also known as sensitivity or true positive rate, represents the proportion of actual positive instances that were correctly identified by the model. It assesses the model's ability to identify all relevant positive cases. Recall is defined as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

4. **F1-Score:** The F1-Score is the harmonic mean of precision and recall, providing a single metric that balances the two. This measure is particularly useful in scenarios with class imbalance, as it accounts for both false positives and false negatives. The F1-Score is defined as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

5. **F2-Score:** The F2-Score is a weighted harmonic mean of precision and recall, with more emphasis placed on recall. Unlike the F1-Score, which gives equal importance to precision and recall, the F2-Score prioritizes minimizing false negatives over false positives. The F2-Score is defined as:

$$F2 - Score = \frac{5 \times Precision \times Recall}{4 \times Precision + Recall}$$

In the context of credit card fraud detection, the performance metrics must be carefully selected to evaluate the model's ability to detect fraudulent transactions accurately while minimizing false positives. The key requirements for these metrics are as follows:

- **High Recall:** Since the primary objective of the model is to detect fraudulent transactions with high accuracy, the recall metric must be prioritized. Recall measures the proportion of true positive results among all actual positive instances. A high recall indicates that the model is capable of identifying most fraudulent transactions, which is crucial in fraud detection.

- **High Precision:** In addition to achieving a high recall, the model should also exhibit high precision. Precision measures the proportion of true positive results among all predictions made by the model that are classified as positive. High precision means the model is effective in predicting fraudulent transactions, thereby reducing the occurrence of false positives, which is essential for minimizing the disruption caused by false alarms.

- **High F2-Score:** The F2-score is a weighted harmonic mean of precision and recall, with greater emphasis placed on recall. This metric is particularly valuable in fraud detection because it ensures that the model prioritizes the identification of true fraudulent transactions, while still considering the importance of precision. A high F2-score reflects a model that achieves a good balance between detecting fraud and minimizing false positives.

In summary, a high-performance fraud detection model enables financial institutions to identify fraudulent transactions in real-time, preventing potential financial losses and enhancing the security of credit card transactions. Achieving high classification performance on imbalanced datasets requires the model to prioritize high recall, precision, and F2-score to effectively address the challenges posed by rare fraudulent events.

The dataset used is widely utilized in academic research for the development and evaluation of machine learning algorithms in the field of fraud detection. For example,

previous studies have leveraged this dataset to explore the efficacy of deep learning algorithms in detecting fraudulent transactions (Phua et al., 2019) and to examine the impact of feature selection on the performance of machine learning models (Ou et al., 2021). Multiple ML techniques will be employed to detect fraudulent credit card transactions. A comparative analysis will then be conducted to evaluate the performance of each technique and identify the most effective model for detecting fraud.

4. Data understanding

In the second step of the CRISP-DM methodology, Data Understanding, the dataset will be described, including an analysis of the variables and their distributions. Table 2 shows the class, number, and value of transactions.

The variable "Class" is represented by the following numbers:

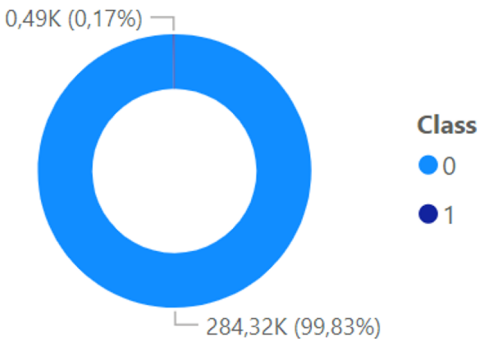
- 0 - which is identified as "non-fraudulent" and
- 1 - which is identified as "fraudulent".

Table 2 *Number and value of non-fraudulent (0) and fraudulent (1) transactions*

Class	Transactions	Amount
0	284315	25.102.462,04 €
1	492	60.127,97 €
Total	**284807**	**25.162.590,01 €**

Figure 1

Percentage of non-fraudulent (0) and fraudulent (1) transactions



Note. Figure 1 clearly shows the class imbalance. The minority class, 0 (fraudulent transactions), accounts for only 0.17% of the total number of transactions.

All variables in the dataset, except for "Class," are shown as decimal numbers (float).

The distribution of the "Class" variable is divided into transactions with class 0 (non-fraudulent), which account for 284,315 instances, and transactions with class 1 (fraudulent), which account for 492 instances. Additionally, the dataset contains no missing values in any column.

The "Time" column represents the number of seconds that have passed between each transaction and the first transaction in the dataset. The "Amount" column represents the transaction amount, while the columns from V1 to V28 are anonymized for data confidentiality, and further details about these variables cannot be provided.

Because the variables from V1 to V28 are anonymized, we can't precisely determine what they represent. However, we can speculate on what these anonymized variables may potentially represent based on general knowledge of credit card fraud detection and typical preprocessing techniques:

- Transaction Information: These variables could represent anonymized transaction details, such as the transaction amount, currency code, or the time of day when the transaction occurred.

- User Behavior Patterns: These might be derived from user behavior patterns, such as transaction frequency, transaction speeds, or time intervals between transactions.

- Geographical Data: Anonymized geographical information, such as the country or region of the transaction.
- Purchase Category: Characteristics that represent the type of purchase or the product category being bought.
- Merchant Information: Data related to the merchant involved in the transaction, possibly anonymized merchant IDs or types.
- Card Information: Anonymized data related to the credit card used in the transaction, such as the card type or issuer information.
- Aggregate Statistics: Aggregate characteristics, such as the average transaction amount or the number of transactions within a specific time frame.

In summary, based on the exploratory analysis, several conclusions were drawn:

- The dataset is highly imbalanced, with a majority of legitimate transactions (99.83%) → data balancing is necessary to adequately train machine learning algorithms.
- All attributes, except for "Time" and "Amount," are normalized around 0 → normalization of the "Time" and "Amount" variables is required.
- The large number of 30 variables makes the training of machine learning models computationally expensive and time-consuming →it is necessary to select the most relevant variables.
- The direct relevance of variables cannot be assessed due to data anonymization → feature selection methods, such as correlation-based variable selection, should be applied.
- There are no missing values in any column →no additional data preprocessing (e.g., interpolation) is needed to fill in missing values.

5. Data preparation

The third phase of the CRISP-DM methodology is Data Preparation. In this phase, data is collected, cleaned, transformed, and integrated to create a structured and reliable dataset suitable for analysis. Key tasks include handling missing values, detecting outliers, selecting relevant variables, and formatting data for modeling. As previously discussed, proper data preparation is essential for the effective training of machine learning algorithms. Data preparation consists of four key steps: data normalization, dataset splitting for training and testing, data balancing, and feature selection. The first step—data normalization—is discussed in detail below.

5.1. Data Normalization

The first preprocessing step involves scaling (normalizing) the variables "Time" and "Amount", along with features V1–V28. Normalization ensures that all variables are on a similar scale, preventing certain features from disproportionately influencing the model due to their magnitude. This step enhances the learning process by ensuring that each variable contributes proportionally to the model training (Han et al., 2011).

5.2. Splitting Data into Training and Test Sets

After normalization, the dataset is split into training and test sets to prevent data leakage and biased results that may arise when machine learning algorithms are trained exclusively on the original dataset. A commonly recommended split ratio is 80:20, ensuring a balanced representation of both legitimate and fraudulent transactions in both subsets (Provost & Fawcett, 2013).

The training set (80% of the original dataset) undergoes further preprocessing, including data balancing, feature selection, and model training. In contrast, the test set (20%) remains untouched and is used solely for model validation. This approach provides an unbiased assessment of algorithm performance on previously unseen data, ensuring a more reliable evaluation of the model's effectiveness. Based on test set validation results, the most suitable algorithm is selected for deployment.

5.3. Data Balancing

As discussed in Section 2, the issue of imbalanced datasets leads machine learning algorithms to classify most transactions as legitimate, effectively ignoring fraudulent transactions due to their low frequency. Additionally, the large volume of transactions slows down the training process of ML algorithms. An imbalanced dataset also hinders

the accurate assessment of which variables are relevant for identifying fraudulent transactions.

To address this, a combined method was used to balance the dataset. First, legitimate transactions were under-sampled, followed by over-sampling of fraudulent transactions.

Under-sampling was performed using the One-Sided Selection (OSS) method, which aims to remove majority-class examples that are far from the decision boundary. Applying OSS helps reduce the number of transactions while preserving representative information about legitimate transactions. After under-sampling, the fraud-to-legitimate transaction ratio remained high at 394:23,664 (approximately 1:60). Generally, imbalanced datasets have ratios ranging from 1:4 to 1:100 (Krawczyk, 2016). Therefore, further processing was required through over-sampling.

Over-sampling of fraudulent transactions was conducted using the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2011). SMOTE generates synthetic samples by selecting data points in feature space, drawing a line between them, and creating new samples along this line. However, it is important to note that generating synthetic transactions does not make the fraudulent sample more representative. Additionally, an excessive number of transactions may slow down ML training. Therefore, a 1:2 fraud-to-legitimate transaction ratio was chosen to avoid excessive synthetic fraud transactions while still balancing the dataset. As a result of the data balancing process, the final training dataset contained 11,832 fraudulent and 23,664 legitimate transactions, achieving a 1:2 ratio. This ensures that the training dataset is no longer highly imbalanced while preventing an overgeneration of synthetic fraudulent transactions.

5.4. Feature Selection

Following data balancing, feature selection was performed using Pearson correlation analysis. Out of the 30 available variables in the training set, the 15 most strongly correlated variables were selected. Feature selection is necessary because some variables are uninformative, and including all variables in training incurs high computational costs. The test dataset was not subjected to feature selection. Instead, the validation of trained models on the test dataset will be performed using the 15 most relevant variables selected from the training set. At this stage, the data has undergone all preprocessing steps and is now ready for model training.

6. Data modeling

The fourth phase of the CRISP-DM methodology is Data Modeling. In this phase, data mining techniques are applied to the prepared dataset to build and train predictive or descriptive models. Various machine learning algorithms are evaluated and selected, and model performance is assessed using techniques such as cross-validation. The primary objective of this phase is to develop efficient models capable of identifying patterns and relationships within the data.

Five machine learning (ML) algorithms were used for modeling. These include one linear algorithm: Logistic Regression, and four non-linear algorithms: Decision Trees, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest. Each of these algorithms has its advantages, but it is not possible to determine in advance which one will yield the best predictive performance.

- Logistic Regression is well-suited for basic linear patterns.
- Decision Trees handle more complex relationships by partitioning data based on feature importance.
- Random Forest enhances performance by using an ensemble of decision trees to improve predictive accuracy.
- KNN focuses on local patterns in the data by considering the nearest neighbors to classify new instances.
- SVM is particularly effective for high-dimensional data, making it a promising choice since the training dataset consists of 15 attributes, making it multidimensional.

Each of the five models was trained on the training dataset using predefined parameters. After training, the models were validated on the test set. The validation results are presented in the following section.

7. Model evaluation

The machine learning models were evaluated on the test dataset, which consists of data not used during training. This test set maintains the same class imbalance as the original dataset: 56,864 (99.83%) legitimate transactions and 98 (0.17%) fraudulent transactions. As discussed in Section 2, the following metrics should be applied when evaluating models on imbalanced datasets:

- Precision
- Recall
- F2-score

The evaluation results are presented in figures 2-4.

Figure 2

Precision results

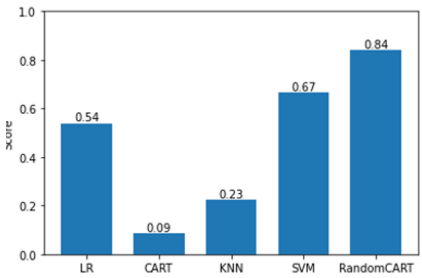


Figure 3

Recall results

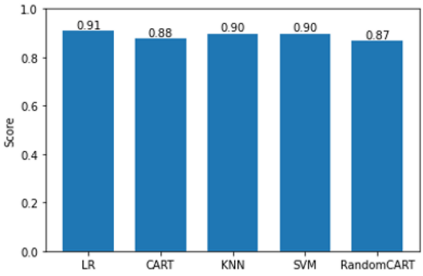
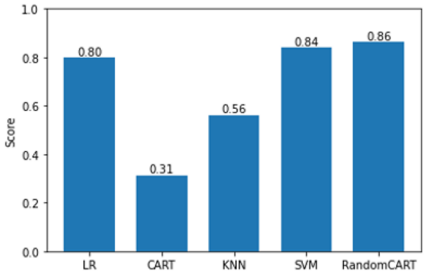


Figure 4

F2-score results



- LR - Logistic Regression
- CART - Classification and Regression Trees (Decision Trees)
- KNN - K-Nearest Neighbors
- SVM - Support Vector Machines
- RandomCART - Random Forest

As shown in the previous three graphs, precision is highest for the SVM and Random Forest models. A low precision value indicates a higher number of false positives, meaning that legitimate transactions were misclassified as fraudulent. This type of error is less critical than failing to detect an actual fraudulent transaction. The recall metric measures the proportion of correctly identified fraudulent transactions relative to the total number of fraudulent transactions. The recall values for the five models are similar, ranging between 0.87 and 0.91. Among the tested models, Logistic Regression

detected the highest number of fraudulent transactions, while Random Forest performed the worst in this regard.

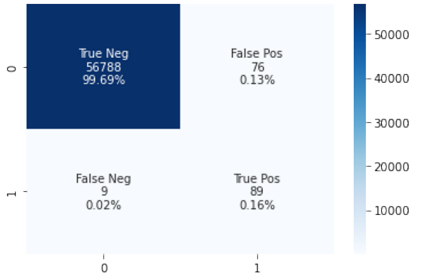
The F2-score is a weighted average of precision and recall, with greater emphasis on recall, as recall is more important than precision in fraud detection scenarios. In this case, the three best-performing models were Logistic Regression, SVM, and Random Forest.

Although these three metrics allow for a qualitative comparison of all five models, they are not sufficient to determine the final model selection. Therefore, a confusion matrix was constructed for each of the five models to provide a quantitative distribution of predictions across the four possible outcomes (see Figure 5).

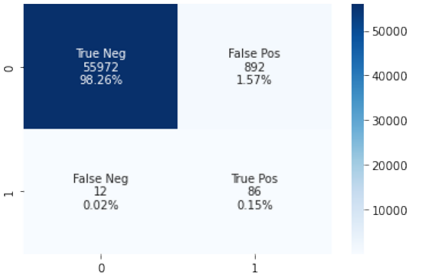
Figure 5

Evaluation of the model on the test dataset based on confusion matrices

a) Logistic Regression

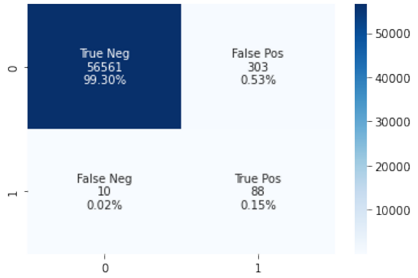


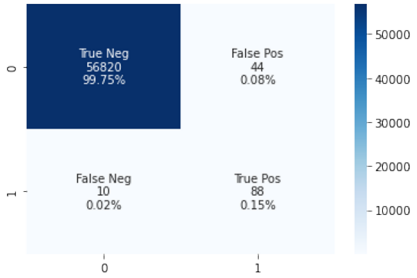
b) Decision Trees



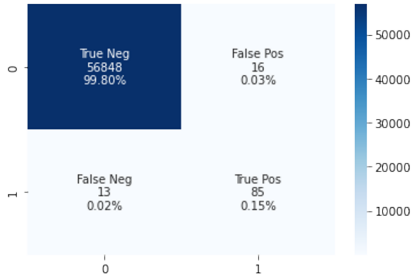
c) KNN

d) SVM





e) Random Forest



The low precision of the Decision Tree and K-Nearest Neighbors (KNN) models is reflected in the expectedly high number of false positives, with 892 and 303 misclassified transactions, respectively. While this number may not seem significant in the context of a total of 56,962 transactions, other models demonstrated better performance with fewer false positives. For instance, the Logistic Regression model has the lowest number of false negatives, which is particularly important from a banking reputation standpoint. The Random Forest model has the fewest false positives among all models but exhibits the highest number of false negatives. The Support Vector Machine (SVM) model represents a balanced trade-off, with only one more false negative than Logistic Regression while maintaining a relatively low number of false positives (44 misclassifications).

In summary, three models—Logistic Regression, SVM, and Random Forest—demonstrated the best performance, confirming the ability of machine learning algorithms to detect complex patterns in large datasets. Based on these results, the

bank's management decides which of the three models to implement and use for real-world data in the deployment phase.

These research findings support Hypothesis H1, confirming the capability of machine learning algorithms to effectively detect financial fraud. The Logistic Regression, SVM, and Random Forest models have proven highly effective in identifying suspicious patterns and anomalies in financial transactions, enabling financial institutions to proactively combat fraudulent activities (Nguyen et al., 2022). Regarding Hypothesis H2, the paper shows that machine learning algorithms can significantly reduce the number of fraudulent transactions by efficiently identifying most of them. While H2 is validated, maintaining high effectiveness requires continuous improvement and optimization to ensure low false positive and false negative rates. To achieve this, machine learning experts can train new, more robust models on updated data, fine-tune training parameters, and test new algorithms.

Finally, the ability of machine learning to facilitate real-time fraud detection (H3) has transformed how financial institutions approach fraud prevention. Instead of reacting to fraud after it occurs, banks can now detect fraudulent activities as they happen and take immediate measures to prevent further damage (Goodfellow et al., 2016). This proactive approach not only enhances the security of financial transactions but also fosters customer trust.

However, as demonstrated in our findings, machine learning models are not yet capable of detecting all fraudulent transactions. Recall scores indicate that despite the application of machine learning algorithms, there remains a risk that some fraudulent activities may go undetected. Additionally, models frequently misclassify legitimate transactions as fraudulent, as shown by precision scores. These results highlight that there is still significant room for improvement in these algorithms, as they currently do not provide entirely accurate fraud detection.

8. Conclusion

This paper conducted a comparative evaluation of five machine learning algorithms—Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Random Forests—for detecting credit card fraud using a highly imbalanced, real-world dataset. Guided by the CRISP-DM methodology, the paper systematically addressed the business objective of identifying fraudulent transactions while minimizing false negatives, which is critical for effective financial fraud prevention.

The findings show that Logistic Regression achieved the highest recall (91%), making it the most effective at detecting fraudulent transactions among the tested models. SVM followed closely with a recall of 89%, offering a balanced performance across all evaluation metrics. Random Forest demonstrated the highest precision (98%), effectively minimizing false positives but at the cost of a higher false-negative rate. KNN and Decision Trees underperformed in both recall and precision, making them less suitable for real-world deployment in this context.

From a practical perspective, the choice of algorithm should depend on institutional priorities. If minimizing false negatives and detecting the maximum number of fraudulent transactions is the highest priority, Logistic Regression is the recommended option. If a more balanced trade-off between recall and precision is required, SVM may be more suitable. Conversely, if reducing false positives is critical—such as when the cost of false alerts is particularly high—Random Forest provides the strongest performance. By clarifying these trade-offs, this study not only contributes to the academic discussion but also offers actionable guidance for financial institutions in selecting the most appropriate algorithm for fraud detection.

By using appropriate performance metrics—recall, precision, and F2-score—the paper ensured that model evaluation was aligned with the risk-sensitive nature of fraud detection, where false negatives carry greater business cost than false positives. The paper also demonstrated the importance of data preparation steps such as normalization, feature selection, and class balancing using SMOTE techniques, which significantly improved model performance.

The key contribution of this research lies in its comparative focus on multiple well-established algorithms, its practical alignment with fraud detection requirements, and its transparent methodology that can be replicated or extended in future work. Unlike many existing studies, this emphasizes the trade-offs between detection accuracy and operational feasibility in fraud detection, offering insights that are directly applicable to financial institutions seeking data-driven solutions.

Nevertheless, the models evaluated are not without limitations. Despite improvements, none achieved perfect detection, and some fraudulent transactions were still missed. Additionally, synthetic oversampling introduces limitations related to generalizability and may not fully represent real fraud patterns. Future work should explore more advanced algorithms, such as ensemble learning with hybrid models or deep learning techniques, and focus on real-time deployment and model retraining on continuously updated data.

As a conclusion, Logistic Regression, SVM, and Random Forest emerged as the most promising algorithms for credit card fraud detection in this paper. Their relative strengths provide financial institutions with viable options depending on whether recall, precision, or a balance of both is the operational priority. While machine learning does not eliminate fraud, it clearly offers a powerful tool for reducing risk and enhancing financial security when applied with careful data preparation and appropriate evaluation standards.

References

1. Ashtiani, M.N., Raahemi, B.: Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review. *IEEE Access* **10**, 72504–72525 (2022). https://doi.org/10.1109/ACCESS.2021.3096799
2. Aslam, F., Hunjra, A.I., Ftiti, Z., Louhichi, W., Shams, T.: Insurance fraud detection: Evidence from artificial intelligence and machine learning. *Research in International Business and Finance* **62**, 101744 (2022). https://doi.org/10.1016/j.ribaf.2022.101744
3. Bin, R., Vitaly, S., Paul, S.: Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems* **2**(1), 55–68 (2022). https://doi.org/10.1007/s44230-022-00004-0
4. Chapman, P., Custer, B., Shearer, C.: CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM Consortium (2000)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357 (2011)
6. Dowling, M., Aziz, S., Hammami, H.: Machine learning in finance: A topic modeling approach (2019). https://doi.org/10.13140/RG.2.2.22110.69447
7. Finance Research Letters: Evaluating Fintech industry's risks: A preliminary analysis based on CRISP-DM framework. *Finance Research Letters* **55**(B), 103966 (2023). https://doi.org/10.1016/j.frl.2023.103966
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
9. Guo, L., Song, R., Wu, J., Xu, Z., Zhao, F.: Integrating a machine learning-driven fraud detection system based on a risk management framework (2024). https://doi.org/10.20944/preprints202406.1756.v1
10. Gupta, P., Shuaib, M., Alam, S., Shuaib, M., Alam, S.: Machine learning-oriented comparative paper of balancing techniques. *Procedia Computer Science* **218**, 2575–2585 (2023). https://doi.org/10.1016/j.procs.2023.01.231
11. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. 3rd edn. Morgan Kaufmann, San Francisco (2011)
12. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284 (2009)
13. Herath, H., Kulasooriya, K.: Credit card fraud detection and prevention strategies. *Journal of Financial Security* **11**(2), 10–15 (2015)

14. Ileberi, E., Sun, Y., Wang, Z.: Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost. *IEEE Access* **9**, 165286–165294 (2021)
15. Ileberi, E., Sun, Y., Wang, Z.: A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data* (2022). https://doi.org/10.1186/s40537-022-00573-8
16. Khemakhem, M.: Evaluating machine learning models for credit card fraud detection: Precision, recall, F1-score, and beyond. *Journal of Computer Science* **11**(4), 285–295 (2020)
17. Krawczyk, B.: Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence* **5**(4), 221–232 (2016)
18. Machine Learning Group – ULB: Credit card fraud detection dataset. Kaggle. Homepage, https://www.kaggle.com/datasets/ulb/creditcardfraud (2019)
19. Mahmood, T., Hashemi, S.K., Mirtaheri, S.L., Greco, S.: Machine learning techniques for detecting fraud in credit card transactions. In: *SEBD 2023*, pp. 469–478 (2023)
20. Mutemi, A., Bacao, F.: E-commerce fraud detection based on machine learning techniques: Systematic literature review. **7**(2) (2024). https://doi.org/10.26599/BDMA.2023.9020023
21. Nguyen, N.G., Pham, H., Bao, B.P., Nguyen, T.: Machine learning approaches for fraud detection in financial transactions: A review. *IEEE Access* **10**, 18423–18437 (2022)
22. Ou, Z., Xie, L., Zhang, W.: The role of feature selection in improving machine learning algorithms for fraud detection. *International Journal of Machine Learning and Cybernetics* **11**(2), 10–22 (2021)
23. Pahuja, L., Kamal, A.: EnLEFD-DM: Ensemble learning based Ethereum fraud detection using CRISP-DM framework. *Expert Systems* **40**(9), e13379 (2023)
24. Phua, C., Lee, V., Smith, K.: Credit card fraud detection using deep learning techniques. *Journal of Financial Technology* **7**(1), 1–15 (2019)
25. Provost, F., Fawcett, T.: Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media, Sebastopol (2013)
26. Ramzan, M., Ahmed, M.: Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access* **10**, 39700–39715 (2022). https://doi.org/10.1109/ACCESS.2022.3166891
27. Rukhsar, L., Bangyal, W.H., Nisar, K., Nisar, S., On, R., On, A.: Prediction of insurance fraud detection using machine learning algorithms. *Mehran University Research Journal of Engineering and Technology* **41**(1), 33–40 (2022)
28. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science* **2**(3), 1–21 (2021). https://doi.org/10.1007/s42979-021-00592-x