Beyond Accuracy: Economic Performance of Machine Learning Models in Financial Fraud Detection

Pedro-Pablo Chambi-Condori * , Miriam Chambi-Vásquez , Telma Saravia-Ticona

Posted Date: 25 February 2026

doi: 10.20944/preprints202602.1609.v1

Keywords: financial fraud detection; operational risk; cost-sensitive learning; expected loss; ap-plication of machine learning in finance

 doi:10.20944/preprints202602.1609.v1

Article

Beyond Accuracy: Economic Performance of Machine Learning Models in Financial Fraud Detection

Pedro-Pablo Chambi-Condori [1,*], Miriam Chambi-Vásquez [2] and Telma Saravia-Ticona [1]

[1] Jorge Basadre Grohmann National University, Avenida Bolognesi / Avenida Pinto s/n, Tacna, Peru

[2] National University of San Marcos, Av. Venezuela Cdra. 34, Lima 1, Peru

* Correspondence: pchambic@unjbg.edu.pe

Abstract

Financial fraud is one of the biggest operational risks for financial institutions, generating significant financial losses and destabilizing the market. While machine learning models are good at predicting, their evaluation often relies on statistical performance metrics that don't directly translate into financial impact. This research develops an evaluation framework that integrates the costs of early fraud detection with predictive effectiveness and economic criteria for decision-making. Several supervised learning models (XGBoost, neural network, random forest, decision tree, and logistic regression) were trained and tested on an unbalanced dataset of credit card transactions. To measure the potential benefit of the models for financial institutions, the savings rate and expected loss were used, along with classic metrics such as F1 score, AUC-PR, AUC-ROC, recall, and accuracy. The economic results are highly sensitive to models with similar predictive capabilities. The ensemble methods, in particular, achieved the optimal balance between fraud detection capabilities and loss reduction, while models optimized solely for accuracy resulted in higher operating costs due to false positives or undetected fraud. The results indicate that the choice of fraud detection models should not be based solely on predictive accuracy, but also on cost asymmetry and risk tolerance. The proposed framework offers practical guidance to financial institutions seeking to align operational risk management and regulatory requirements with the implementation of machine learning, enabling risk-informed decision-making.

Keywords: financial fraud detection; operational risk; cost-sensitive learning; expected loss; application of machine learning in finance

1. Introduction

The economic stability of companies, consumers, and financial institutions is threatened by financial fraud (Mustafa et al., 2024). The gradual digitization of financial services has increased the number and speed of transactions, while also expanding opportunities for fraudulent activities that exploit both human and technological weaknesses. According to global industry estimates, financial fraud results in annual losses of billions of dollars worldwide (Alloy, 2024). For known fraud patterns, traditional detection methods, which rely primarily on manual audits and pre-established rules, are useful; however, they cannot detect complex, changing, and unusual behaviors. Therefore, due to its ability to learn from large volumes of transactional data, detect hidden patterns, and adapt to new behaviors in near real time, machine learning (ML) has become an essential tool for early fraud detection.

Recent studies confirm that the use of machine learning techniques for fraud detection is on the rise. As we have seen, algorithms are widely used to detect financial anomalies, particularly credit card fraud. Examples of such algorithms include support vector machines and neural networks (Abdulalem, 2022). Depending on the use case and risk profile, different machine learning algorithms demonstrate complementary strengths (Choi and Gipper, 2024; Riskiyadi, 2025; Su et al., 2025). However, most studies evaluate models primarily using technical metrics such as AUC-ROC or

 doi:10.20944/preprints202602.1609.v1

accuracy, which can be misleading in highly unbalanced datasets, common in financial transaction data. Although false negatives are more costly than false positives, little research has been done on the economic consequences of misclassification. Thus, this study proposes a comparative cost analysis of different machine learning models. This paper uses expected loss and accuracy-recovery analysis as decision criteria to support operational risk management in real-world fraud detection systems.

Recent comparative studies show that machine learning algorithms are not always effective in detecting financial fraud. For example, Random Forest outperforms other models in training and test datasets, while logistic regression and support vector machines are consistently reliable across different scenarios (Lee et al., 2025). According to Zhao and Bai (2022), combined and hybrid approaches, such as XGBoost and logistic regression, have also achieved high predictive accuracy in detecting corporate fraud (>97%). While these results demonstrate good technical performance, they are generally measured using overall accuracy metrics, which can mask operational shortcomings when the fraud rate is very low. In real-world scenarios, if false negatives remain frequent, a high-accuracy model can lead to financial losses.

Fraud detection is related to credit risk management, where financial losses occur when debtors default on their contractual obligations (Bhatore et al., 2020). Predictive modeling is increasingly being applied by financial institutions to monitor behavioral patterns, prevent fraud, and manage non-performing assets. Hybrid and neural network-based models are being used more and more frequently for default prediction, anomaly detection, and credit scoring, according to several literature reviews. However, these techniques primarily improve statistical predictive capacity rather than the economic impact of decisions. Thus, the gap between operational profitability and predictive accuracy persists, demonstrating the need for evaluation frameworks that consider the asymmetry of financial costs in the selection and implementation of models.

According to Hilpisch (2020), artificial intelligence has advanced rapidly in the last 20 years thanks to access to data, connectivity, and increased computing power. Traditional rating agencies have been gradually complemented by machine learning in finance, which offers automated and scalable methods for risk assessment. Supervised learning algorithms have been effective in classifying financial risks, with empirical evidence from emerging markets. A large study in Saudi Arabia, for example, found that K-Nearest Neighbors, Gradient Boosting, Decision Tree, and Random Forest algorithms achieved high predictive capacity in risk classification (Alsuwailem et al., 2023). In annual cross-sectional tests, Random Forest was always correct in more than 90% of cases, consistently outperforming alternative classifiers.

However, while these results are promising, most empirical studies focus on technical performance measures, such as accuracy, precision, AUC-ROC, F1 score, and recall. While these measures inform statistical discrimination capabilities, they don't always reveal how effective operations are in situations with large class imbalances and unequal misclassification costs. In real-world financial systems, a small improvement in recovery can translate into large monetary savings, and high overall accuracy can coexist with significant losses from undetected fraud. Therefore, there is still a gap between predictive optimization and cost-effective implementation. This reinforces the need for cost-sensitive evaluation frameworks that explicitly incorporate expected monetary loss when comparing models.

The limitations of conventional statistical and rule-based methods for detecting fraud in financial statements have also been addressed in recent research. However, these conventional approaches often overlook subtle, nonlinear patterns indicative of fraudulent behavior (Hossain et al., 2024). To fill these gaps, machine learning models such as Random Forest, XGBoost, and support vector machines have been applied to large datasets containing financial ratios, governance measures, and company-specific characteristics. Predictive capability has been further enhanced through advanced preprocessing strategies, including oversampling, imputation of missing values, and feature scaling. Experimental results consistently demonstrate that the combined approaches outperform conventional statistical models in terms of recall, accuracy, and AUC-ROC.

 doi:10.20944/preprints202602.1609.v1

Furthermore, fraud models should incorporate non-financial governance measures, such as board independence and audit fees (Cheng-Wen et al., 2025). These findings suggest that the problem of fraud detection is not just about finding misplaced numbers, but about structurally analyzing governance. However, these studies demonstrate that machine learning is becoming increasingly effective in audit and regulatory contexts, yet it continues to prioritize statistical validation metrics. Little attention has been paid to evaluating how model performance translates into economic benefit, considering the costs of asymmetric misclassification. Therefore, integrative frameworks that link model prediction with the effectiveness of economic decision-making remain necessary.

According to Olowe et al. (2024), predictive modeling and machine learning (ML) technologies are transforming financial services. These technologies enable financial institutions to leverage big data, improve decision-making, risk management, and the customer experience. This report explores how machine learning and predictive modeling are transforming financial services in areas such as fraud detection, customer scoring, market forecasting, and risk management. Predictive modeling techniques, such as regression, decision trees, and support vector machines, are essential for optimizing portfolios, assessing credit risk, and improving financial projections. At the same time, ML algorithms—specifically deep learning and natural language processing—can detect fraud, transaction anomalies, and extract insights from unstructured data, such as financial reports or social media. The report also illustrates how ML is being applied to customer interactions and how fintechs and banks are developing personalized services and products that drive customer loyalty.

However, despite these potential benefits, challenges remain in implementing these technologies, particularly regarding data privacy, model interpretability, and regulatory compliance. With the digitalization of the financial world, it is necessary to combat algorithms and protect information. This example demonstrates how ethics and innovation must work together to generate transparency and trust. To quickly recap existing applications, industry use cases, and future trends, predictive analytics and machine learning (ML) are transforming the future of financial services. "It opens the door to ongoing research and collaboration to discover new possibilities and build a sustainable sector." According to Valavan et al. (2022), machine learning (ML) algorithms are the backbone for identifying credit and debit card fraud and loan defaults. These algorithms can be trained on past fraud data or historical data, allowing them to recognize patterns in current or future transactions. As with most datasets, fraudulent transactions are far rarer than legitimate ones, making fraud detection difficult. The best way to prevent loan defaults is to detect delinquent loans early. Machine learning algorithms are increasingly performing better with this data due to their computational power.

Common classification metrics, such as recall, precision, F1 score, ROC curve, and accuracy (Kumar et al., 2020), have been used to compare different machine learning algorithms for credit card fraud detection. These algorithms include decision trees, neural networks, support vector machines, logistic regression, and gradient boosting. While these approaches achieve high predictive power, fraud datasets are often highly unbalanced, reducing the reliability of accuracy-based assessments and their practical interpretability in the financial world. In the real world, undetected fraudulent transactions can generate much greater losses than legitimate transactions that are mislabeled. This underscores the need for evaluation metrics aligned with financial impact, not just statistical performance.

Therefore, this paper compares different machine learning models for the early detection of financial fraud using a cost-sensitive evaluation framework. Accuracy-recovery and expected loss metrics are used as the preferred metrics for evaluating the models: logistic regression, decision trees, random forests, XGBoost, and neural networks.

The goal is to determine which model provides the best balance between predictive power and operational cost efficiency for highly unbalanced transactional data.

 doi:10.20944/preprints202602.1609.v1

1.1. Análisis de la Literature

1.1.1. Financial Fraud in the Digital Economy

Financial fraud is one of the greatest threats to the stability of the global financial system. The digitization of financial services, electronic payments, and online commerce has made operations more efficient. But it has also opened the door to more fraud (Ngai et al., 2011). Today, financial criminals are employing more sophisticated methods that combine social engineering, automation, and the exploitation of technological vulnerabilities, creating constantly evolving patterns that are difficult to detect with traditional means.

Fraud is adversarial in nature, whereas other financial risks, such as market or credit risks, are not. Fraudsters continually adapt their methods to avoid detection by existing systems, thus generating what is known as concept change in transactional data (Bolton & Hand, 2002). Therefore, detection systems must be adaptive and capable of learning emerging patterns in near real time.

Fraud affects the economy. Institutions also face indirect costs in the form of reimbursements, regulatory fines, reputational damage, and investigations, in addition to the immediate financial losses from unauthorized transactions (Button et al., 2014). Therefore, early fraud detection has become an essential part of corporate risk management.

Traditionally, financial institutions have used rules based on transaction limits or unusual locations. However, these approaches have significant limitations, as they fail to identify subtle anomalies or complex patterns (Phua et al., 2010). The inability of these methods to adapt to new forms of fraud has opened the door to data-driven machine learning techniques.

1.1.2. Machine Learning for Fraud Detection

The main analytical tool for financial fraud is machine learning, as it can model nonlinear relationships in high-dimensional data (Dal Pozzolo et al., 2015; Gotelaere & Paoli (2025)) and Chen et al. (2025). Conventional statistical methods are more complex, since transactional databases contain many implicit variables related to user behavior.

Logistic regression is one of the most widely used and highly regarded supervised algorithms due to its stability and interpretability (Bahnsen et al., 2016). Decision trees can capture interactions between variables. On the other hand, ensemble methods, such as Random Forest, reduce variance by averaging many models (Breiman, 2001). According to Chen & Guestrin (2016), gradient-powered algorithms, such as XGBoost, can iteratively correct prediction errors with high accuracy.

Hossain et al. (2024) have also successfully used artificial neural networks, especially in cases with complex nonlinear patterns. However, their use in finance faces regulatory barriers due to their lack of interpretability.

According to Dal Pozzolo et al. (2015), many comparative studies have shown that ensemble methods outperform individual models in classification metrics such as the F1 score and ROC-AUC. But fraud detection involves judgments with real-world financial consequences (Bahnsen et al., 2014), and greater predictive capacity does not always translate into better operational performance.

1.1.3. The Issue of Class Imbalance

High skewness is a key characteristic of datasets related to financial fraud. Less than 1% of all observations are fraudulent transactions, which introduces significant bias into model training (He and Garcia, 2009).

In this scenario, accuracy ceases to be a relevant evaluation criterion. A model that considers all transactions legal could achieve accuracy values higher than 99% without detecting any fraud. For this reason, alternative metrics focused on the minority class are used.

Accuracy assesses the number of positive predictions that are correct, while recall indicates the model's ability to identify genuine frauds. The F1 metric, according to Saito and Rehmsmeier (2015), combines both measurements into a balance between false positives and false negatives.

 doi:10.20944/preprints202602.1609.v1

The area under the ROC curve (ROC-AUC) has traditionally been used to analyze discriminatory potential. However, this metric could be overly optimistic in data with a highly uneven balance, considering the true negative rate (Davis and Goadrich, 2006). Therefore, for fraud problems, the precision-recall (PR-AUC) curve is considered more appropriate.

To mitigate the effects of imbalance, preprocessing methods such as synthetic data generation using SMOTE (Chawla et al., 2002), oversampling, and undersampling are employed. These techniques enable models to recognize significant patterns without bias toward the predominant class.

1.1.4. Evaluation Based on Economic Costs

Statistical metrics allow for the comparison of models, but they do not demonstrate their usefulness in real-world financial situations. Errors in fraud detection have unequal costs: a false negative means a direct financial loss, while a false positive causes customer inconvenience and operational expenses (Bahnsen et al., 2016).

Cost-based learning incorporates these economic effects when evaluating the model (Elkan, 2001). Instead of reducing classification errors, the goal is to decrease the expected economic loss.

The Expected Loss indicator calculates the total economic impact of model errors, assigning a monetary value to false negatives and false positives (Bahnsen et al., 2014). Furthermore, the Savings Rate evaluates the reduction in losses compared to a scenario without a detection system. Although of paramount practical importance, most studies continue to prioritize conventional metrics over economic indicators (Ngai et al., 2011). This gap limits the possibility of translating academic findings into practical business applications.

1.1.5. Financial Interpretation and Regulations

Financial institutions operate under rigorous regulations that require automated decisions to be transparent. Validating complex models, such as deep neural networks, is complicated because their interpretation presents challenges; this hinders their validation in regulatory audits (Rudin, 2019).

To address this problem, explainable artificial intelligence (XAI) methods, such as LIME and SHAP, have been developed, which make it possible to determine the contribution of each variable to the prediction (Lundberg & Lee, 2017). These tools allow the adoption of advanced models without jeopardizing the traceability of decisions.

Therefore, there is a trade-off between predictive capacity and interpretability. Simple models are more transparent but less efficient, while complex models are more predictable, although less clear.

1.1.6. Research Gaps

The literature shows that machine learning has enabled significant progress in improving detection accuracy. However, three fundamental limitations remain: Statistical metrics prevail over economic ones. There is a lack of evaluation frameworks focused on operational decision-making. Comparative analysis of models considering the true financial impact is scarce.

Therefore, financial institutions lack standardized methods for selecting models according to risk tolerance and business objectives.

2. Materials and Methods

2.1. Research Design

The effectiveness of different machine learning models for the early detection of financial fraud in an economic risk scenario is analyzed in this study, which uses an experimental and quantitative methodology. Cost-based evaluation and predictive performance analysis are combined to align the model evaluation with operational risk management standards.

 doi:10.20944/preprints202602.1609.v1

Unlike traditional comparative research that focuses solely on statistical discrimination metrics, this study incorporates expected loss and savings indicators to analyze the financial consequences of classification-related decisions. This perspective is based on the cost-sensitive learning framework presented by Elkan (2001) and implemented by Bahnsen et al. (2014, 2016) in financial fraud situations.

The methodology comprises five key phases: (1) dataset selection, (2) data preprocessing, (3) model creation, (4) training and validation, and (5) performance evaluation in statistical and economic terms.

2.2. Explanation of the Dataset

The empirical analysis uses the publicly available credit card fraud identification dataset, which was initially published on Kaggle. This dataset is widely used in academic research (Dal Pozzolo et al., 2015). The database consists of 284,807 credit card transactions made in September 2013 by European cardholders, of which 492 (0.172%) were identified as fraudulent.

All input variables, except for Amount and Time, are principal components derived from the transformation of anonymous features. The extreme class imbalance of the dataset makes it suitable for analyzing fraud detection systems under realistic operating conditions (He and Garcia, 2009).

The dependent variable is binary, where:

Y=1 if the transaction is fraudulent; otherwise, Y=0.

2.3. Data Preprocessing

Given the significant imbalance between legitimate and fraudulent transactions, data preprocessing plays a crucial role in the model's effectiveness (Chawla et al., 2002).

The following steps were implemented:

- Attribute Scaling:

To ensure numerical stability in models sensitive to magnitude disparities, the Quantity variable was normalized using Min-Max scaling.

- Feature Selection:

The Time variable was not included in the modeling process due to its minimal contribution to predictive capacity and the impossibility of interpreting the transformed feature space.

- Class Imbalance Control:

Two resampling methods were implemented:

  - Random subsampling of the predominant class: Synthetic Minority Oversampling (SMOTE) method to produce synthetic fraud observations (Chawla et al., 2002).
  - Data partitioning: To maintain class proportions, the dataset was partitioned into training (80%) and test (20%) subsets using stratified sampling.

To prevent data leakage, all preprocessing procedures were implemented only on the training set.

2.4. Computer Learning Models

This research analyzes five supervised classification algorithms frequently used in the fraud detection literature.

2.4.1. Logistic Regression

The logit link function allows logistic regression to calculate the probability of fraud:

$$P(Y = 1) = \frac{1}{1 + e^{-(B_0 + B_1 X_1 + B_2 X_2)}}$$

Where P(y=1) indicates the probability of the event of interest occurring. Predictor variables used to calculate the probability are X_1, X_2, …, X_n. Coefficients B_0, B_1,…,B_n are calculated from the data and indicate the change in the logit probability for each unit of variation in the relevant variable. Logistic regression is used as a reference model due to its interoperability and the probabilistic nature of its results (Bhnsen et al., 2016).

 doi:10.20944/preprints202602.1609.v1

2.4.2. Decision Trees

Decision trees repeatedly divide the feature space to minimize impurity measures, such as the Gini index. While they offer interpretability, they can suffer from overfitting in high-dimensional contexts.

According to Kotsiantis (2013), the construction of a decision tree begins with all the data at the root. The algorithm then chooses the feature that best separates the data into two or more sets at each stage. This procedure is repeated for each new node until the stopping condition is met. Decision trees are used in classification and regression. Root node: the initial point where the first division is made. Internal nodes: intermediate decisions that divide the information. Leaves (final nodes): final conclusions or predictions.

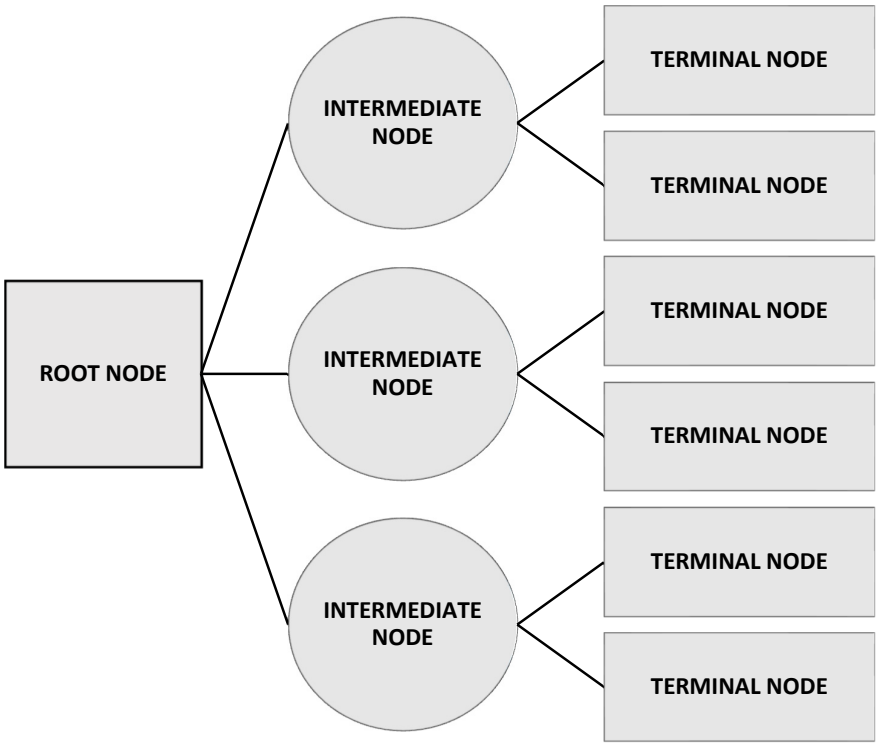

Figure 1. Illustration of the Decision Tree Algorithm. *Note*. Original work.

2.4.3. Random Forest

The Random Forest constructs multiple decision trees using bootstrap sampling and feature randomization, aggregating predictions through majority voting (Breiman, 2001). This ensemble method reduces variance and improves generalization.

According to López (2018) and Parmar et al. (2019), the Random Forest approach combines several random selection trees and sums their predictions using an average. Its high accuracy and superior performance have been recognized. It is an ensemble model based on the combination of multiple decision trees, used for classification tasks.

2.4.4. Gradient Boosting (XGBoost)

Gradient boosting builds models sequentially, correcting previous errors. XGBoost introduces regularization to avoid overfitting and improve computational efficiency (Chen and Guestrin, 2016).

According to Dorogush et al. (2018), the Gradient Boosting technique is a supervised learning algorithm that builds sequential models with the goal of correcting errors in the previous model. It is mainly used for classification and regression problems and is known for its ability to generate highly

 doi:10.20944/preprints202602.1609.v1

accurate models. Its foundation is gradient descent, which allows it to iteratively optimize a loss function through a series of weak models.

FO(X), with which the boosting algorithm begins, defined as follows:

$$F_o(X) = argmin_y \sum_{i=1}^{n} L(Y_i, ¥)$$

The gradient of this function is calculated iteratively:

$r_{im} = -\alpha \left[\partial \frac{(L(Y_i, F(X_i)}{\partial F(X_i)}\right] F(X) = F_{m-1}(X)$, where α is the learning rate.

2.4.5. Artificial Neural Network

A forward propagation neural network with hidden layers was used to capture complex and nonlinear relationships. Neural networks, despite their power, are difficult to interpret in regulated financial environments (Rudin, 2019).

Neural networks, according to Mishra & Srivastava (2014), are machine learning models based on the human brain. They consist of artificial neurons arranged in layers, whose objective is to detect complex patterns in data to perform tasks such as classification, regression, and others. They are powerful and flexible, but require a large volume of data and high computing power. They have achieved significant progress in areas such as computer vision, natural language processing, and recommendation systems due to their ability to learn complex representations.

All models were trained with the same processed data, with predetermined hyperparameters tuned by cross-validation (k-fold with k=5) and hyperparameter fitting with Grid Search or Bayesian Optimization, depending on the model.

2.5. Model Training and Validation

Hyperparameter fitting was performed using grid search with cross-validation to optimize model performance and reduce the risk of overfitting. Five-step cross-validation was applied to the training set to ensure robust parameter estimation.

Model calibration was also considered to improve the reliability of the probability estimation, following the recommendations of Bahnsen et al. (2016).

2.6. Performance Evaluation Metrics

2.6.1. Statistical Metrics

Predictive performance was evaluated using the metrics of accuracy, precision, recallability, F1 score, AUc-ROC, and AUC-PR. While these metrics provide information on discriminatory power, they do not consider the economic consequences of misclassification (Saito and Rehmsmeier, 2015).

2.6.2. Economic Evaluation

In this section, to integrate the financial impact, two economic metrics were implemented:

First, Expected Loss (EL).

This was calculated to incorporate operational relevance, and a cost-sensitive evaluation was introduced using the expected loss, expressed by the following equation:

$$E(L) = FN * C_{FN} + FP * C_{FP}$$

Where C_FN denotes the cost of an undetected fraudulent transaction and C_FP is the cost of incorrectly blocking a legitimate transaction. This framework follows cost-sensitive evaluation principles (Elkan, 2001; Bahnsen et al., 2014). This metric quantifies the financial impact of misclassification errors and allows for comparison between models beyond statistical accuracy.

Second, the Savings Rate (SR)

Savings were calculated relative to a baseline scenario without fraud detection:

$$\text{SR} = \frac{\text{Loss}_{\text{baseline}} - \text{Loss}_{\text{model}}}{\text{Loss}_{\text{baseline}}}$$

 doi:10.20944/preprints202602.1609.v1

This metric quantifies the financial benefit generated by the predictive system.

2.7. Decision Threshold Optimization

Instead of adopting the default probability threshold of 0.5, optimal cut-off points were determined by minimizing expected loss. This approach aligns classification decisions with institutional risk tolerance and cost asymmetry (Bahnsen et al., 2016).

2.8. Robustness and Overfit Control

To ensure the model's generalizability, cross-validation was applied, regularization parameters were adjusted, performance was evaluated using unanalyzed test data, and overfitting was monitored by comparing training and test performance.

2.9. Methodological Contribution

The methodological contribution of this study lies in the integration of: predictive modeling with imbalance detection, hyperparameter optimization with cross-validation, cost-sensitive evaluation based on expected financial loss, and threshold optimization aligned with risk management objectives. This framework allows for the selection of models with financial information, reducing the gap between machine learning performance metrics and operational risk decision-making.

2.10. Evaluation Metrics

Given the significant class imbalance, metrics particularly sensitive to this type of problem were used:

- Precision
- Sensitivity or Recall
- F1-Score
- AUC-ROC (Area Under the Receiver Operating Characteristic Curve)
- AUC-PR (Area Under the Precision-Recall Curve)
- False Positive Rate (FPR) and False Negative Rate (FNR)
- CONFUSION MATRIX

	Actually positive	**Actually negative**
Predicted positive	**TP**	**FP**
Predicted negative	**FN**	**TN**

METRICS: Accuracy, Precision, and Recall

True Positive Rate:

$$\text{Recall (or TPR)} = \frac{correctly\ classified\ actual\ positives}{all\ actual\ positives} = \frac{TP}{TP + FN}$$

False Positive Rate:

$$\text{FPR} = \frac{incorrectly\ classified\ actual\ negatives}{all\ actual\ negatives} = \frac{FP}{FP + TN}$$

Precision:

$$\text{Precisión} = \frac{correctly\ classified\ actual\ positives}{everything\ classified\ as\ positive} = \frac{TP}{TP + FP}$$

These metrics allow us to assess the balance between effective fraud detection and minimizing errors that negatively impact the customer (such as unjustified account blocks).

2.11. Validation and Comparability

To ensure comparability and reproducibility, all models were trained and evaluated under the same sampling and computational conditions. Testing was conducted in a controlled environment

 doi:10.20944/preprints202602.1609.v1

using Python and libraries such as TensorFlow/Keras for autoencoders and neural networks, LightGBM, imbalanced-learn, scikit-learn, and XGBoost.

3. Results

The database comprises 284,807 transactions, each with 30 attributes and a target variable called "Class." No values are missing from this set. However, there is a significant difference, as only 0.17% of the transactions are fraudulent. StandardScaler was used to normalize the Amount and Time characteristics. The data were then divided into two sets: a training set (80%) and a test set (20%), ensuring that the class distribution remained consistent through stratification.

A balanced class weight logistic regression model was used for training. This model correctly identified 92% of the fraudulent transactions, resulting in an impressive recovery rate of 0.92 for the fraud class. However, the model had quite low accuracy (0.06) in the fraud category, indicating that a large number of legitimate transactions were misclassified as fraudulent, producing 1,410 false positives. Although the overall accuracy was high (0.98), this metric is not as valuable when working with highly unbalanced datasets. The ability to differentiate between the two classes is exceptional, as the AUC ROC score was 0.97. Furthermore, as can be seen in Figure 2, the logistic regression model performed robustly with a high AUC score (0.97), similar to that of the neural network.

Although recall and accuracy are high, the extraordinarily low precision (FP = 1410) indicates a significant number of false alarms. Therefore, the model classifies an excessive number of legal transactions as fraud, resulting in high operating costs for auditing. This demonstrates that precision and recall alone are insufficient for evaluating anti-fraud models. The model exhibits typical behavior for linear classifiers on unbalanced datasets: it prioritizes recall over accuracy. It detects virtually all frauds; however, it does so with a high number of false positives. This makes it operationally impractical.

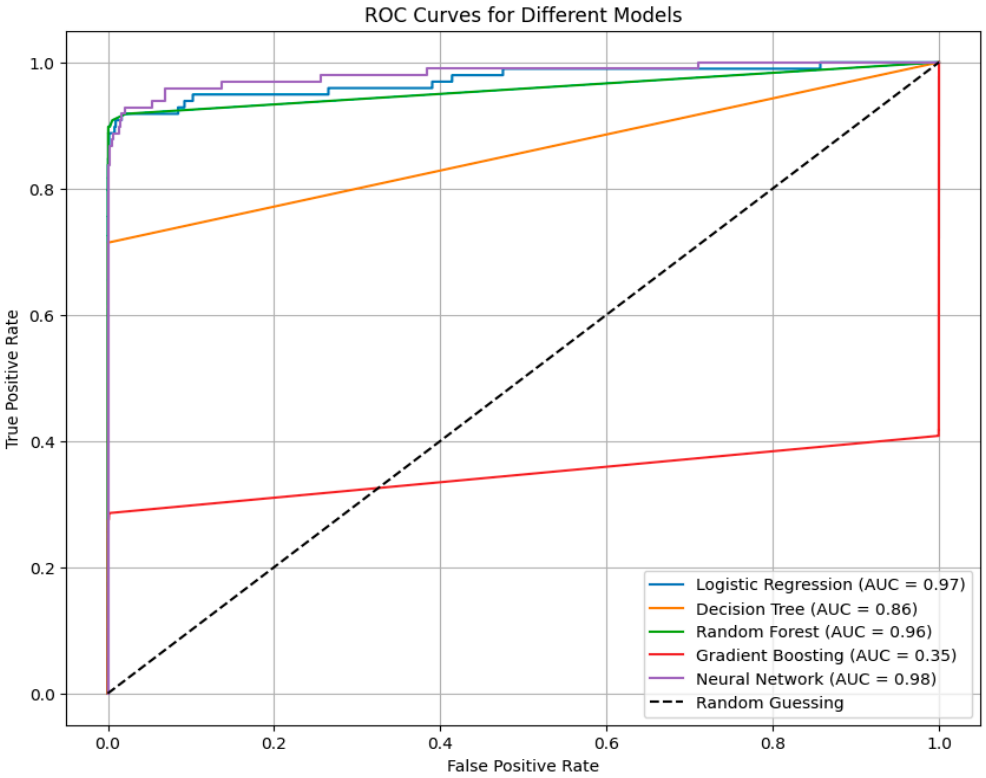

Figure 2. ROC curves for the five evaluated models. *Note.* Prepared by the authors using sample data and support from Google Colab.

 doi:10.20944/preprints202602.1609.v1

Table 1. Indicators of accuracy of the Logistic Regression algorithm.

Confusion Matrix				
	Precision	**Recall**	**f1-Score**	**Support**
0	1.00	0.98	0.99	56,864
1	0.06	0.92	0.11	98
Accuracy			0.98	56,962
Macro avg	0.63	0.95	0.55	56,962
Weighted avg	1.00	0.98	0.99	56,962
ROC AUC score: 0.97				

The neural network model exhibits the highest AUC score (0.98), as shown in Table 2 and Figure 2, indicating that it has the best overall ability to distinguish between fraudulent and non-fraudulent transactions among the tested models. Its ROC curve is closest to the upper left corner. The model is highly selective, only flagging fraud when it is very confident. This reduces operating costs but increases losses due to undetected fraud.

Table 2. Neural Network Model Accuracy Indicators.

Confusion Matrix				
	Precision	**Recall**	**f1-Score**	**Support**
0	1.00	1.00	1.00	56,864
1	0.96	0.55	0.70	98
Accuracy			1.00	56,962
Macro avg	0.98	0.78	0.85	56,962
Weighted avg	1.00	1.00	1.00	56,962
ROC AUC score: 0.98				

The random forest model exhibits a good AUC score (0.96), with good performance, although slightly inferior to that of the neural network and logistic regression in terms of overall discriminatory capacity, as measured by AUC. The model maintains high accuracy and good recall simultaneously. It generates only 3 false positives and detects most frauds. This indicates that ensemble methods reduce the overfitting typical of individual trees. When comparatively evaluating the AUC+PR value, it is the model that best identifies the minority class (Fraud).

Table 3. Accuracy indicators of the Random Forest Model Classification Report.

Confusion Matrix				
	Precision	**Recall**	**f1-Score**	**Support**
0	1.00	1.00	1.00	56,864
1	0.96	0.76	0.85	98
Accuracy			1.00	56,962
Macro avg	0.98	0.88	0.92	56,962
Weighted avg	1.00	1.00	1.00	56,962
ROC AUC score: 0.96				

The decision tree model exhibits a lower AUC score (0.856) compared to logistic regression, random forest, and neural network, suggesting it is less effective at distinguishing between classes. The model significantly improves accuracy compared to the logistic regression model, drastically reducing false positives; however, it loses its ability to detect fraud (FN=28).

 doi:10.20944/preprints202602.1609.v1

Table 4. Decision Tree Model Accuracy Indicators.

Confusion Matrix				
	Precision	**Recall**	**f1-Score**	**Support**
0	1.00	1.00	1.00	56,864
1	0.71	0.71	0.71	98
Accuracy			1.00	56,962
Macro avg	0.85	0.86	0.86	56,962
Weighted avg	1.00	1.00	1.00	56,962
ROC AUC score: 0.86				

The gradient boosting model exhibits the lowest AUC score (0.35) (see Table 5), indicating poor performance on this task. Its ROC curve is near or below the random guess line, suggesting that its performance is not significantly better than chance. In summary, based on the ROC analysis, neural network and logistic regression models appear to be the most promising for this fraud detection task, closely followed by the random forest model. The decision tree and, especially, the gradient boosting models did not perform as well according to this metric. The model does not detect fraud (FN=80). This may be due to a conservative decision threshold or because the model prioritizes minimizing positive frauds, which is inappropriate in an anti-fraud context.

Table 5. Gradient Model Accuracy Indicators Boosting.

Confusion Matrix				
	Precision	**Recall**	**f1-Score**	**Support**
0	1.00	1.00	1.00	56,864
1	0.53	0.18	0.27	98
Accuracy			1.00	56,962
Macro avg	0.76	0.59	0.64	56,962
Weighted avg	1.00	1.00	1.00	56,962
ROC AUC score: 0.35				

A comparative analysis of the results was performed, using the selected metrics as a reference. Topics discussed included:

- Overall model performance.
- Ability to detect fraud (recall).
- Resilience to unbalanced data.
- Interpretation of the model in the context of finance.

It is crucial to keep in mind that, although AUC is a good general measure, other metrics such as recall and accuracy should also be considered, especially when working with unbalanced datasets and the specific requirements of a fraud detection system (such as the cost of false positives relative to false negatives). We have already discussed these metrics. The recall-accuracy curve shows the variation in recall and accuracy at different classification thresholds. Accuracy: This is the model's ability to avoid classifying something as positive when it is actually negative. In the context of fraud detection, this equates to the proportion of transactions that are identified as fraudulent and actually are (true positives / (true positives + false positives)). High accuracy implies a decrease in false alarms. On the other hand, recall measures the model's ability to identify all positive samples. When it comes to fraud detection, this is understood as the proportion of fraudulent transactions that are accurately identified (true positives / (true positives + false negatives)). A high recovery rate means fewer frauds go undetected. The curve starts in the upper left corner of Figure 2, where recovery is almost zero and accuracy is high (or undefined). As we move to the right along the curve, recovery increases, although accuracy generally decreases. The single value that summarizes performance across all possible thresholds is the area under the curve (AUC) of the accuracy-recovery curve (PR AUC). A

 doi:10.20944/preprints202602.1609.v1

higher PR AUC indicates better performance, particularly on unbalanced datasets, where the ROC AUC could be misleading. For our logistic regression, the PR AUC is 0.97.

The trade-off can be seen from the example thresholds below the graph: When the threshold is very low, the model tends to predict almost everything as positive, resulting in high recovery (1.0000) but fairly low accuracy (0.0017). This means it identifies all fraud, but also categorizes almost all legitimate transactions as fraudulent. Accuracy improves (to 0.0056) as the threshold increases (for example, to 0.0593), although recovery begins to decline (to 0.9592). The model becomes more selective in its fraud predictions, which reduces false positives; however, it may still miss some fraudulent transactions. With a higher threshold (for example, 0.1679), recovery drops again to 0.9490, and accuracy increases further (to 0.0135). We observe that, with the example threshold of 0.8, accuracy increases significantly to 0.16, while recall remains high at 0.89. This demonstrates that we can considerably reduce the number of false positives (from 1386 at the default threshold to 458 at 0.8) while identifying a significant portion of fraudulent transactions (87 out of 98) by increasing the threshold. Key takeaway: The accuracy- recall curve allows us to observe and understand the inherent trade-off between detecting all fraudulent transactions (high recall) and reducing false alarms (high accuracy). Selecting the ideal threshold depends on the specific priorities and costs of your anti-fraud system. Developing an accurate machine learning model to identify fraud is essential, but its true impact is achieved when integrated with existing business procedures and systems. Below, we explain what enterprise integration entails and why it is so crucial.

Alignment with business goals: The model must be aligned with the company's goals. For example, would it prioritize reducing financial losses from fraud (which implies greater recovery) or decreasing the operating cost of investigating false positives (which implies greater accuracy)? The chosen model and its threshold must be consistent with these priorities. Define the prediction as "applicable": When the model predicts "fraud," this must have a specific meaning for the company. Some options would be: - Stop the transaction immediately. - Flag the transaction for a fraud analyst to review manually. - Ask the customer to confirm the transaction. - Change the customer's risk assessment. Workflow integration: The model's results must be optimally integrated into the existing fraud detection process. This could mean: - Receiving real-time transaction data. - Sending the model's predictions to a fraud management system.

Enable tasks or alerts for the fraud investigation team. Establishing the threshold based on a cost-benefit analysis: For classification purposes, the ideal threshold should not rely solely on model performance metrics. It should also consider the operational and financial costs of false positives and negatives. A cost-benefit analysis can be helpful in identifying the point at which total costs are minimized or the company's net benefits are maximized. Feedback loop: To achieve continuous improvement, it is crucial to establish a feedback loop. This means: - Gathering feedback from fraud analysts on the model's predictions, such as whether a detected transaction was truly fraudulent. - Using this feedback to retrain and update the model frequently.

Fraud detection systems, from a legal and compliance standpoint, must adhere to data privacy laws and regulations. You must verify that the model and its implementation meet these criteria during integration. Furthermore, for successful integration and to ensure all stakeholders understand and accept the model's strengths and weaknesses, it's essential to engage with all involved parties (IT teams, fraud analysts, risk managers, compliance officers, etc.). In other words, contextualizing the business means grounding the data science solution in reality. Verify that the model is not only a technological advancement but also a tool for preventing and reducing financial fraud within the organization.

 doi:10.20944/preprints202602.1609.v1

Table 6. Comparative evaluation of fraud detection models and economic considerations.

Model	Accuracy	Recall	Precision	F1-Score	AUC-ROC	AUC-PR	FN	FP	Expected Loss (USD)	Savings Rate (%)
Logistic Regression	0.98	0.92	0.06	0.11	0.97	0.055	8	1410	6,630	77.50
Decision Tree	1	0.71	0.71	0.71	0.86	0.260	28	28	2,828	71.10
Random Forest	1	0.76	0.96	0.85	0.96	0.360	23	3	6,909	76.50
XGBoost	1	0.18	0.53	0.27	0.35	0.048	80	16	8,016	18.00
Neural Network	0.98	0.55	0.96	0.7	0.98	0.265	44	2	13,206	55.10

Note. Prepared by the author using sample data.

4. Discussion

Comparative analysis validates that machine learning models offer efficient methods for early detection of financial fraud, particularly in contexts where fraudulent behavior patterns are dynamic and nonlinear. Logistic regression showed consistent performance across all evaluation metrics, while random forests and neural networks demonstrated strong predictive ability, consistent with previous research (Abdulalem, 2022; Lee et al., 2025; Zhao and Bai, 2022). However, the current findings extend previous research by showing that predictive performance alone is not a determinant of operational utility.

This study makes a crucial contribution by identifying discrepancies between statistical evaluation metrics and economic performance. Models with high recovery rates, such as neural networks and logistic regression, optimized fraud detection coverage; however, they produced many false positives, which could increase the operational burden. In contrast, while random forests provided a balanced predictive profile, they did not always reduce economic losses. Notably, the model with the lowest expected loss was not the same one that achieved the best predictive metrics, demonstrating that common comparisons based on accuracy can lead to poorer implementation decisions in financial real-world scenarios.

These results support the recent literature that argues fraud should be analyzed as a cost-sensitive decision problem rather than a classification task. Including expected loss as an evaluative standard offers a more accurate representation of operational performance, since financial institutions are asymmetrically impacted by classification failures. False negatives cost money in lost sales, while false positives cost in operational efficiency and customer experience. Therefore, the choice of model should be determined not only by predictive metrics but also by institutional risk appetite.

The findings also highlight the importance of balancing model complexity with its practical applicability. Ensemble models and neural networks generally achieved greater predictive power, but at the cost of more computational resources and less transparency. Interpretability remains crucial in regulated financial applications, requiring the incorporation of interpretability techniques such as SHAP or LIME to support the auditing and regulatory compliance of high-performance models.

Overall, the literature shows that combining economic evaluation criteria with machine learning allows for better alignment between operational risk management and predictive analytics. Future research should address the adaptive optimization of thresholds and cost-effective online learning policies to enable on-the-fly adjustments to detection policies in response to evolving fraud patterns.

 doi:10.20944/preprints202602.1609.v1

5. Conclusions

This research addressed the detection of financial fraud using machine learning models trained on highly unbalanced transactional data. The results show that traditional performance metrics (AUC-ROC, accuracy) are insufficient to determine the true effectiveness of detection systems. This is because models with similar predictive capabilities generated very different economic and operational impacts.

The comparative analysis showed that logistic regression achieved the best recall rate, but with a high false positive rate that limits its practical use. The neural network performed with high accuracy, but low recall, missing many fraudulent cases. Statistically, the Random Forest model offered the best balance between accuracy and recall. However, the Decision Tree model resulted in the lowest expected financial loss, proving that the model with the best predictive performance does not always maximize monetary gains!

These findings confirm that the choice of fraud detection models must be based on cost-sensitive assessments, not just conventional statistical metrics. Including expected loss as a decision criterion significantly alters the categorization of models and leads to a more realistic approach to operational performance, enabling the detection system to align with business objectives and risk tolerance.

To minimize financial losses while maintaining operational efficiency and customer experience, future research directions should focus on methods for optimizing cost-dependent thresholds and frameworks for continuously updating models.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, P.-P. C.-C.; Methodology, P.-P. C.-C.; Software, P.-P. C.-C.; Investigation, M. C-V.; Data curation, M. C-V.; Writing—review and editing, T. S.-T.; Visualization, T. S.-T.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: These data were derived from the following resources available in the public domain: Kaggle, https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

Conflicts of Interest: The authors declare no conflict of interest.

References

Abdulalem, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences*. https://doi.org/10.3390/app12199637

Alloy (2024). State of Fraud Benchmark Report. https://www.alloy.com/state-of-fraud-benchmark-report-2024-ty

Alsuwailem, A.A.S., Salem, E. & Saudagar, A.K.J. (2023). Performance of Different Machine Learning Algorithms in Detecting Financial Fraud. *Comput Econ 62*, 1631–1667. https://doi.org/10.1007/s10614-022-10314-x

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2014). Example-dependent cost-sensitive logistic regression for credit card fraud detection.

Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2016). Improving credit card fraud detection with calibrated probabilities.

Bhatore, S., Mohan, L. & Reddy, Y.R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *J BANK FINANC TECHNOL 4*, 111–138. https://doi.org/10.1007/s42786-020-00020-3

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, *17*(3), 235–249. http://www.jstor.org/stable/3182781

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

 doi:10.20944/preprints202602.1609.v1

Button, M., Lewis, C., & Tapley, J. (2014). Not a victimless crime: The impact of fraud on individual victims and their families. *Security Journal, 27*, 36–54. https://psycnet.apa.org/doi/10.1057/sj.2012.11

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. https://doi.org/10.1145/2939672.2939785

Chen, Y., Zhao, C., Xu, Y., Nie, C., Zhang, Y. (2025). Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications, *Data Science and Management*, https://doi.org/10.1016/j.dsm.2025.08.002

Cheng-Wen L., Mao-Wen, F., Chin-Chuan, W, & Muh. I. (2025). Evaluating Machine Learning Algorithms for Financial Fraud Detection: Insights from Indonesia. *Mathematics.* https://doi.org/10.3390/math13040600

Choi, J. & Gipper, B. (2024). Fraudulent financial reporting and the consequences for employees, *Journal of Accounting and Economics. Volume 78*, Issue 1. https://doi.org/10.1016/j.jacceco.2024.101673

Dal Pozzolo, A., Caelen, O., Johnson, R., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. https://doi.org/10.1109/ssci.2015.33

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. https://dl.acm.org/doi/epdf/10.1145/1143844.1143874

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *ArXiv.* https://arxiv.org/abs/1810.11363

Elkan, C. (2001). The foundations of cost-sensitive learning. https://www.researchgate.net/publication/2365611_The_Foundations_of_Cost-Sensitive_Learning

Gotelaere, S., Paoli, L. (2025). Prevention and Control of Financial Fraud: a Scoping Review. Eur J Crim Policy Res 31, 1–21 (2025). https://doi.org/10.1007/s10610-022-09532-8

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://ieeexplore.ieee.org/document/5128907

Hilpisch, Y. J. (2020). Artificial intelligence in finance : a Python-based guide (First edition.). *O'Reilly Media*, Inc. https://go.oreilly.com/stanford-university/library/view/-/9781492055426/?ar

Hossain, M.Z., Raja, M.R., & Hasan, L. (2024). Developing Predictive Models for Detecting Financial Statement Fraud: A Machine Learning Approach. European Journal of Theoretical and Applied Sciences, 2(6), 271-290. DOI: https://doi.org/10.59324/ejtas.2024.2(6).22

Kotsiantis, Sotiris. (2013). Decision trees: A recent overview. *Artificial Intelligence Review.* 1-23. https://doi.org/10.1007/s10462-011-9272-4.

Kumar, N. , Simaiya, S. , Lilhore, U. & Kumar S. An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods. *International Journal of Advanced Science and Technology Vol. 29*, No. 5, (2020), pp. 3414 – 3424. ISSN: 2005-4238 IJAST.

Lee, C.-W., Fu, M.-W., Wang, C.-C., & Azis, M. I. (2025). Evaluating Machine Learning Algorithms for Financial Fraud Detection: Insights from Indonesia. *Mathematics, 13*(4), 600. https://doi.org/10.3390/math13040600

López, M. (2018). Advances in Financial Machine learning.Willey. ISBN 978-119-43211-6. https://agorism.dev/book/finance/ml/Marcos%20Lopez%20de%20Prado%20-%20Advances%20in%20Financial%20Machine%20Learning-Wiley%20%282018%29.p.. https://www.rockwellautomation.com/content/dam/rockwell-automation/documents/pdf/campaigns/state-of-smart-2025/INFO-BR027D-ES-P-noi.pdf df

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. https://doi.org/10.48550/arXiv.1705.07874

Mishra, M. & Srivastava, M. (2014). "A view of Artificial Neural Network", International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), Unnao, India, 2014, pp. 1-3, https://doi.org/10.1109/ICAETR.2014.7012785

Mustafa Abdul Salam, Khaled M. Fouad, Doaa L. Elbably, Salah M. Elsayed. Federated learning model for credit card fraud detection with data balancing techniques. *Neural Comput & Applic 36*, 6231–6256 (2024). https://doi.org/10.1007/s00521-023-09410-2

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). Data mining for financial fraud detection. Decision Support Systems, 50(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

 doi:10.20944/preprints202602.1609.v1

Olowe, K., Edoh N.,Zouo, S. & Olamijuwon. J. (2024). Review of predictive modeling and machine learning applications in financial service analysis. *Computer Science & IT Research Journal*, 5(11), 2609-2626. https://doi.org/10.51594/csitrj.v5i11.1731

Parmar, A., Katariya, R., Patel, V. (2019). A Review on Random Forest: An Ensemble Classifier. In: Hemanth, J., Fernando, X., Lafata, P., Baig, Z. (eds) International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 26. *Springer,* Cham. https://doi.org/10.1007/978-3-030-03146-6_86

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. https://doi.org/10.1109/ICICTA.2010.831

Riskiyadi M (2025), "Detecting financial statement fraud using new ensemble learning: evidence during the COVID-19 pandemic in Indonesia". *Journal of Financial Crime*, Vol. 32 No. 4 pp. 825–842, doi: https://doi.org/10.1108/JFC-08-2024-0264

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions. https://doi.org/10.1038/s42256-019-0048-x

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Su, H., Jiang, I & Liu, D. (2025). Detecting financial fraud risk using machine learning: Evidence based on different categories and matching samples, *Finance Research Letters*, Volume 85, Part A. https://doi.org/10.1016/j.frl.2025.107858

Valavan, M. & Rita, S. (2022). Predictive-Analysis-based Machine Learning Model for Fraud Detection with Boosting Classifiers M. Valavan and S. Rita. *Computer Systems Science & Engineering* DOI: https://doi.org/10.32604/csse.2023.026508

Zhao, Z., & Bai, T. (2022). Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms. *Entropy, 24*(8), 1157. https://doi.org/10.3390/e24081157