REVISTA ESMAT

Article

THE ROLE OF AI IN FRAUD DETECTION AND PREVENTION IN ONLINE PAYMENT SYSTEMS

O papel da inteligência artificial na detecção e prevenção de fraudes em sistemas de pagamentos online

El papel de la inteligencia artificial en la detección y prevención del fraude en los sistemas de pago en línea

Stephen Salve
Doctoral Student in Business Administration (DBA), Western State University (WSU), California, EUA.
https://orcid.org/0009-0005-2420-2955
E-mail: stevensalve@gmail.com

Histórico do artigo
Recebido em: 12 de junho de 2026
Aceito em: 13 de agosto de 2026
Revisado em: 20 de agosto de 2026
Publicado em: 10 de setembro de 2026

Editor responsável: Dr. Tarsis Barreto Oliveira
Editor assistente: Dr. Leonardo de Andrade Carneiro

ABSTRACT

The rapid expansion of digital payment systems has been accompanied by increasingly sophisticated financial fraud, including synthetic identity fraud and account takeover, while rule-based detection remains largely reactive and poorly suited to non-linear, high-dimensional transaction data. This study evaluates and compares the performance of supervised machine learning models for fraud detection in online payment systems. A quantitative, experimental design was applied to the publicly available Credit Card Fraud Detection dataset released by the Machine Learning Group of the Université Libre de Bruxelles in collaboration with Worldline, comprising 284,807 European card transactions recorded over two days in September 2013, of which 492 (0.172%) are fraudulent. The feature set consists of twenty-eight variables obtained by principal component analysis, together with transaction time and amount. The data were partitioned 70:30 with stratification, and the Synthetic Minority Over-sampling Technique was applied to the training partition only. Random Forest, XGBoost and a feed-forward Neural Network were trained under

identical preprocessing conditions and evaluated using precision, recall, F1-score, accuracy and AUC-ROC. XGBoost produced the strongest results, with precision of 0.92, recall of 0.88, F1-score of 0.90 and AUC-ROC of 0.99, followed by the Neural Network and Random Forest. The findings indicate that ensemble gradient boosting offers the most favourable balance between fraud capture and false-alarm cost on severely imbalanced payment data. Inference latency and a live rule-based baseline were not measured and remain necessary steps before operational deployment.

Keywords: Artificial intelligence. Fraud detection. Online payment systems. Machine learning. Class imbalance.

RESUMO

A rápida expansão dos sistemas de pagamentos digitais tem sido acompanhada por fraudes financeiras cada vez mais sofisticadas, enquanto a detecção baseada em regras permanece reativa e pouco adequada a dados transacionais não lineares e de alta dimensionalidade. Este estudo avalia e compara o desempenho de modelos supervisionados de aprendizado de máquina para a detecção de fraudes em sistemas de pagamentos online. Adotou-se delineamento quantitativo e experimental, aplicado ao conjunto de dados público Credit Card Fraud Detection, do Machine Learning Group da Université Libre de Bruxelles e da Worldline, com 284.807 transações com cartão de crédito de portadores europeus registradas em setembro de 2013, das quais 492 (0,172%) são fraudulentas. Os atributos reúnem vinte e oito variáveis obtidas por análise de componentes principais, além do tempo e do valor da transação. Os dados foram particionados em 70:30 de forma estratificada, e a técnica SMOTE foi aplicada apenas à partição de treinamento. Random Forest, XGBoost e uma rede neural feedforward foram treinados sob idênticas condições de pré-processamento e avaliados por precisão, recall, F1-score, acurácia e AUC-ROC. O XGBoost apresentou o melhor desempenho, com precisão de 0,92, recall de 0,88, F1-score de 0,90 e AUC-ROC de 0,99. Conclui-se que o gradient boosting por ensemble oferece o equilíbrio mais favorável entre captura de fraude e custo de alarmes falsos em dados severamente desbalanceados. A latência de inferência e uma linha de base baseada em regras não foram mensuradas e permanecem etapas necessárias antes da implantação operacional.

Palavras-chave: Inteligência artificial. Detecção de fraudes. Sistemas de pagamentos online. Aprendizado de máquina. Desbalanceamento de classes.

1 INTRODUCTION

Digital transaction technology has enabled a level of consumer convenience and global economic expansion not previously experienced. However, this increased reliance on technology has also expanded the opportunities available for sophisticated forms of fraud.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

This article examines the contribution of artificial intelligence (AI) to securing online payment systems against continuously evolving threats. It first sets out the shortcomings of traditional methods of detecting and preventing fraudulent activity in the context of modern financial crime, and then examines how machine learning techniques can provide an adaptive and forward-looking response to the continual advancement of financial fraud. The article is organised as follows: Section 1 contextualises the research object, presents the background of the study, states the scientific problem and sets out the objectives; Section 2 constitutes the development of the study, comprising the theoretical framework (subsection 2.1), the research methodology (subsection 2.2) and the discussion of results (subsection 2.3); and Section 3 presents the final considerations, the limitations of the study and directions for future research (Mytnyk *et al.,* 2023).

1.1 Evolution of Digital Transactions and Payment Systems

The advancements in digital transaction technology have brought about dramatic changes in the way financial transactions are made around the world. In the initial stages of electronic payment solutions, provision was limited to basic online banking services and card payments, but with significant improvements in internet connectivity, mobile computing and cloud-based infrastructure, complex digital payment ecosystems have been developed. The current payment landscape comprises mobile payments, real-time payments, contactless payments, embedded finance and international e-commerce platforms (Desai; Kosse; Sharples, 2025).

Although these innovations have made substantial contributions to efficiency, accessibility and economic scalability, they have also widened the attack surface available to offenders. As the volume, velocity and diversity of digital transactions continue to rise, more points of exposure become available to would-be attackers. Fraudulent practices that were once isolated and ad hoc, confined to a single platform, are now frequently organised and automated to exploit systemic weaknesses across multiple platforms (Odufisan; Abhulimen; Ogunti, 2025).

1.2 Role of Artificial Intelligence in Modern Fraud Detection

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

Artificial intelligence has emerged as a strong response to the intrinsic deficiencies of traditional fraud detection methods. Systems based on machine learning allow continuous learning from transaction data and dynamic adaptation to evolving fraud patterns. Their capacity to handle high-dimensional feature spaces supports the identification of faint behavioural anomalies and interdependencies that rule-based mechanisms do not detect (Odufisan; Abhulimen; Ogunti, 2025).

Ensemble methods and neural networks compute probabilistic risk scores for transactions, enabling financial institutions to intervene before fraudulent activity causes significant losses. Unlike static systems, AI-based models update their decision boundaries as new data arrive and therefore remain effective against emerging fraud strategies. This represents a shift from reactive to proactive fraud detection in fintech security and places artificial intelligence at the core of contemporary digital payment protection frameworks (Wang *et al*., 2024).

1.3 Fraud Detection in an Evolving Financial Environment

Fraud detection is the process of identifying and preventing attempts to obtain money or property by deceptive means. It comprises the set of activities undertaken to detect and block such attempts before loss is realised. Fraud detection is applied across banking, insurance, healthcare, government and public administration, as well as in law enforcement.

Within banking, the function has had to adapt as services have moved from branch networks to digital and online platforms, since the controls available in a face-to-face setting do not transfer to channels in which the institution never observes the customer directly (Balcıoğlu, 2024).

1.4 Machine Learning Algorithms and Fintech Security

The financial technology sector has changed how financial services are accessed, delivered and consumed, extending banking functions to mobile applications, embedded payment interfaces and cloud-hosted infrastructure. That extension has moved a substantial share of transaction authorisation away from controlled branch environments and into channels the institution does not own end to end. Security in this setting therefore depends

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

less on perimeter controls than on the ability to assess each transaction on its own characteristics at the moment it is presented, which is the function automated detection is required to perform (Awotunde *et al.,* 2021).

1.5 Limitations of Traditional Fraud Detection Approaches

Traditional fraud control procedures have relied extensively on rule-based reasoning and predefined threshold values. Such procedures performed adequately under earlier transaction volumes but have become inefficient in managing the complexity of contemporary digital payment platforms. Rules are inflexible and cannot represent the non-linear, high-dimensional relationships observed in large-scale financial transaction datasets (Al-Dahasi *et al.,* 2025).

One of the most significant disadvantages of traditional methods is their reactive nature: fraudulent practices are usually identified only after financial damage has occurred. Furthermore, the high false-positive rates produced by rule-based approaches mean that genuine transactions are wrongly flagged as suspicious, degrading customer experience and increasing manual review costs. These challenges are compounded by the class imbalance problem commonly found in fraud data, where only a very small fraction of transactions is fraudulent (Vanini *et al.,* 2023).

1.6 Model Selection Criteria for Fraud Detection Systems

Machine learning algorithms are central to contemporary fraud detection in fintech settings, a position supported by systematic review of the literature on transaction security (Masud et al., 2024). Supervised algorithms such as tree-based ensembles and neural networks are widely applied because they handle large datasets effectively and identify complex, non-linear interactions among features. Their capacity to accommodate the high dimensionality and severe class imbalance that characterise transaction data has been demonstrated empirically across multiple classifiers (Al-Dahasi *et al*., 2025).

Beyond raw detection accuracy, models applied in the financial technology industry must balance precision and recall while remaining computationally efficient. A third criterion applies in regulated settings: supervisory authorities increasingly expect that an automated

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

decision affecting a customer can be explained, which favours models whose outputs can be attributed to identifiable features (Weber; Carl; Hinz, 2024). Systematic selection, training and comparison of machine learning models is therefore essential to the development of robust fraud detection systems. The present work compares multiple machine learning models in order to determine which is best suited to digital payment fraud detection under severe class imbalance (Ravichandran; Inaganti; Muppalaneni, 2023).

1.7 Background of the Study

For financial institutions the consequences of this shift are operational as well as financial. Direct fraud losses are only part of the cost: every legitimate transaction wrongly blocked carries a customer-experience penalty, and every alert raised consumes analyst review time. Institutions therefore require detection methods that can be tuned to a deliberate balance between these two costs, rather than methods that simply maximise the number of transactions flagged. It is against this background that supervised machine learning is examined here — not as a replacement for existing controls, but as a means of setting that balance on an empirical basis (Vanini *et al.,* 2023).

1.8 Problem Statement

While the use of digital payments has increased rapidly in recent years, so too has the incidence of more advanced forms of fraud, such as synthetic identity theft and account takeover. Because traditional rule-based detection systems are primarily reactive, they cannot effectively address these threats or respond to them dynamically. A considerable security gap therefore exists that exposes financial systems to loss. The core difficulty is that existing detection methodologies are not well suited to analysing complex, high-dimensional transaction data or to identifying emerging forms of fraud proactively. This study responds to the need for AI-based methodologies that can learn from data and adjust as fraud patterns change. The scientific problem addressed is therefore the following: which supervised machine learning model provides the most favourable balance between fraud capture and false-alarm cost when applied to online payment transaction data characterised by extreme class imbalance?

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

1.9 Research Objectives

The general objective of this research is to evaluate the effectiveness of supervised machine learning models in detecting fraudulent transactions in online payment systems under conditions of severe class imbalance. The specific objectives are:

1. To examine the application of artificial intelligence techniques to the identification of online payment fraud.
2. To compare the performance of Random Forest, XGBoost and a feed-forward Neural Network on a common, publicly available payment fraud dataset using a single preprocessing and evaluation protocol.
3. To apply synthetic minority over-sampling to the training partition and document its role in enabling the three models to be trained on financial transaction data characterised by extreme class imbalance.
4. To identify, on the basis of the comparative evidence obtained, which of the evaluated models offers the most favourable balance between fraud capture and false-alarm cost, and to specify the conditions that would need to be verified before operational deployment.

2. DEVELOPMENT

2.1 Theoretical Framework

Narender and Anand (2025) examined the role of artificial intelligence in transforming fraud detection and prevention within financial services. The growing intricacy of financial delinquency has rendered conventional identification methods insufficient, necessitating the integration of advanced technologies. The authors present AI as a disruptive force that strengthens defences against fraudulent activity through anomaly detection, predictive analytics and machine learning algorithms. They review several AI approaches used in fraud prevention, including supervised and unsupervised learning, deep learning and natural language processing, and argue that artificial intelligence and big data analytics are essential instruments for mitigating financial crime through enhanced detection precision and faster response times. They further address the collaborative efforts of regulators, technology

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

corporations and financial institutions to establish an ecosystem capable of adapting to shifting fraud tactics.

Odufisan, Abhulimen and Ogunti (2025) argued that fraud represents a substantial risk to Nigeria's emerging digital economy, affecting banking, e-commerce, healthcare and education, and that conventional techniques fail to adapt to advancing fraud tactics. Their article examined the capabilities of artificial intelligence and machine learning for improved fraud detection and prevention in that context, investigating supervised, unsupervised and deep learning approaches and their applications in anomaly detection, behavioural analysis, risk assessment and network analysis. The authors emphasise the advantages of AI-driven fraud detection, including enhanced efficiency, superior accuracy and anticipatory risk management, while recognising technological constraints and regulatory factors as obstacles.

Al-Dahasi *et al.* (2025) examined how the rapid evolution of the internet and of digital payments has transformed financial transactions, producing both technical innovation and a concerning increase in cybercrime. Their paper addresses the challenge of detecting financial fraud in the digital payment landscape and the enhancement of operational risk frameworks. Six machine learning models were applied to fraud detection and their hyperparameters refined, with performance evaluated using accuracy, precision, recall and F1-score. The findings indicate that XGBoost and Random Forest surpass the other models, attaining an equilibrium between false positives and false negatives. The study underscores the need to balance these two error types while guaranteeing precision, scalability, flexibility and transparency.

Wang *et al*. (2024) examined how the rapid advancement of internet banking has exacerbated financial fraud, posing significant difficulties for the security and stability of the financial sector. The study applies artificial intelligence to enhance data mining models for the development of a fraud detection and prevention system. The article sets out the system architecture and the implementation of its key technologies, including model training and optimisation and real-time monitoring and early warning. Analysis of the experimental findings confirms the efficacy and practicality of the system on real financial data, offering a technological assurance and risk management technique for the financial sector.

Balcıoğlu (2024) argued that the incorporation of artificial intelligence and machine

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

learning into risk management and fraud detection in the financial industry represents a notable advance in combating financial crime and enhancing operational efficiency. The chapter investigates the influence of these technologies through case studies of applications in banking and insurance, and examines emerging developments such as federated learning, quantum computing and blockchain for their potential to transform risk management. The author notes that the use of these technologies presents intricate problems and ethical dilemmas relating to data privacy, regulatory adherence and responsible AI use, and underscores the need for stringent data security protocols, transparent algorithms and flexible regulatory frameworks.

Vanini *et al.* (2023) examined online banking fraud, which occurs when a perpetrator gains unauthorised access to accounts and transfers funds from an individual's online bank account. Effectively combating this requires identifying numerous offenders while minimising false positives, which is difficult for machine learning because of significant data imbalance and the intricacy of fraud. The authors delineate three complementary approaches: machine learning-based fraud detection, economic optimisation of machine learning outcomes, and a risk model that forecasts fraud risk while accounting for countermeasures. The models were evaluated on actual data. Their machine learning technique decreased both anticipated and unanticipated losses across three consolidated payment channels by 15% relative to a benchmark of static if-then rules, and further optimisation reduced anticipated losses by 52%, with a false-positive rate of 0.4%.

In summary, current research unequivocally demonstrates that artificial intelligence and machine learning techniques significantly surpass traditional rule-based fraud detection systems regarding accuracy and loss reduction efficacy.

Table 1- Summary of Key Literature on AI-Based Fraud Detection

Author(s) & Year	Context / Domain	Methods / Technologies Used	Key Findings	Identified Gaps / Limitations
Narender and Anand (2025)	Financial services fraud prevention	Supervised & unsupervised ML, Deep Learning, NLP, Big Data Analytics	AI improves fraud detection accuracy, response time and adaptability; supports ethical compliance and regulatory alignment	Lacks empirical comparison using real-world transaction datasets; limited focus on performance metrics under class imbalance

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

Odufisan, Abhulimen and Ogunti (2025)	Nigeria's digital economy (banking, e-commerce, healthcare, education)	Supervised ML, Unsupervised ML, Deep Learning, Behavioural and network analysis.	AI enables proactive fraud detection and adaptive learning against evolving fraud strategies	Context-specific to Nigeria; limited real-time deployment evaluation and comparative algorithm analysis
Al-Dahasi et al. (2025)	Digital payment fraud detection	Six supervised ML models with optimised hyperparameters, including Random Forest and XGBoost; preprocessing by feature selection, standardisation and resampling	XGBoost and Random Forest achieved superior balance between precision, recall, and F1-score	Does not address real-time implementation constraints; interpretability and scalability aspects are limited
Wang et al. (2024)	Internet banking fraud detection	AI-based data mining models, real-time monitoring systems	AI-enhanced systems demonstrate strong effectiveness and practical applicability on real financial data	Focuses more on system architecture than comparative model benchmarking under imbalance
Balcıoğlu (2024)	Banking and insurance risk management	AI, ML, Federated Learning, Blockchain, Quantum Computing	AI and ML enhance fraud detection and operational efficiency; highlights emerging trends	Primarily conceptual and case-study based; lacks quantitative model evaluation
Vanini et al. (2023)	Online banking and payment channels	ML-based fraud detection, risk modelling, economic optimisation	ML reduces expected and unexpected losses by 15% relative to a static rule-based benchmark, rising to 52% after optimisation, with a false-positive rate of 0.4%	Emphasises economic optimisation; limited comparison of multiple ML classifiers under consistent conditions

Source: Prepared by the author based on the reviewed literature (2025). Note: AI = artificial intelligence; ML = machine learning; NLP = natural language processing.

Table 1 summarises the six studies reviewed above. In summary, the current body of research indicates that artificial intelligence and machine learning techniques outperform traditional rule-based fraud detection systems in terms of accuracy and loss reduction. Ensemble learning models such as Random Forest and XGBoost have demonstrated consistently strong performance, while deep learning models show promise in capturing intricate transaction behaviours. Nonetheless, three gaps remain. First, much of the existing work relies on context-specific or restricted datasets, which limits comparability across studies. Second, systematic comparisons of multiple classifiers under a single preprocessing

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

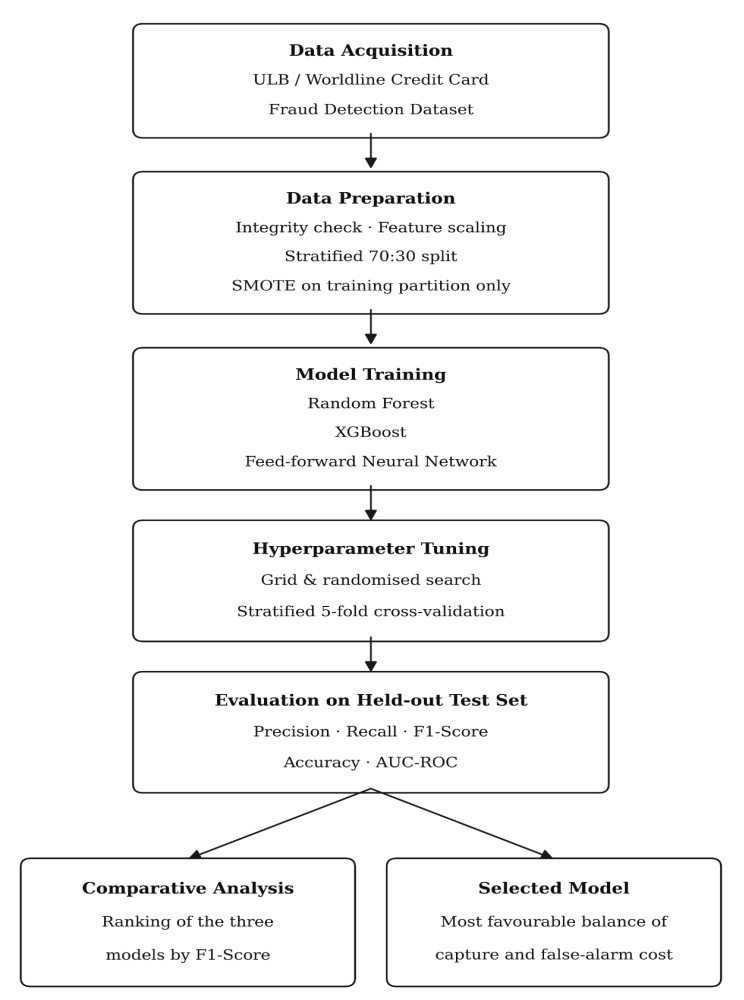
and evaluation protocol in conditions of extreme class imbalance are scarce. Third, claims regarding real-time deployment are frequently made without accompanying measurement of inference latency or throughput. The present research addresses the first two gaps by evaluating three advanced machine learning models on a common public benchmark under identical experimental conditions. The third gap is acknowledged as outside the scope of this study and is treated explicitly in the limitations and future research sections.

2.2 Methodology

A quantitative research design was adopted in order to examine systematically the applicability of machine learning algorithms to fraud detection in online payment systems. A quantitative design is appropriate because the study relies on transaction attributes, statistical results and performance comparisons between algorithms, all of which are objective quantities. The study is applied in nature, explanatory in its objective, and experimental in its procedure. The empirical scope is documentary and secondary: it consists of an anonymised, publicly available transaction dataset.

Figure 1 – Research Design and Analytical Workflow

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360



Source: Prepared by the author (2025).

The interpretive criterion adopted for model comparison is the F1-score on the fraudulent class, complemented by the area under the receiver operating characteristic curve (AUC-ROC) as a threshold-independent measure of ranking quality; accuracy is reported for completeness but is not used as a decision criterion, given that it is uninformative under extreme class imbalance. The overall research design and analytical workflow are summarised in Figure 1.

2.2.1 Research Design

The study is quantitative and employs an experimental, comparative modelling design. Several machine learning models are developed, trained on a common training partition and

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

evaluated on a common held-out test partition using a single set of criteria. This design permits an objective comparison of models because the only variable that changes between experimental conditions is the learning algorithm itself; the dataset, the preprocessing pipeline, the partitioning, the resampling procedure and the evaluation metrics are held constant. Comparative modelling designs of this kind are established practice in fraud detection research within the financial sector (Al-Dahasi *et al*., 2025).

2.2.2 Population of the Study

The target population of this research comprises all digital financial transactions, whether legitimate or fraudulent, processed through online payment platforms. Such transactions are characterised by attributes including transaction value, timestamp, merchant and channel characteristics, and customer behavioural patterns. Because complete transaction records held by financial institutions are commercially sensitive and not available for academic use, the study employs a publicly available, anonymised dataset of real card transactions. This dataset is a bounded sample of that population rather than a representative one, and the consequences for generalisation are set out in the limitations.

2.2.3 Dataset, Source and Sampling

The study uses the Credit Card Fraud Detection dataset compiled during a research collaboration between Worldline and the Machine Learning Group of the Université Libre de Bruxelles (ULB) on big data mining and fraud detection, and made publicly available through the Kaggle platform under the Database Contents License (DbCL) v1.0. (Machine Learning Group - ULB, 2018; Dal Pozzolo *et al*., 2015). The dataset records transactions made with credit cards by European cardholders over two consecutive days in September 2013. It contains 284,807 transactions, of which 492 are labelled as fraudulent, corresponding to 0.172% of all records and a class ratio of approximately 1:578. This degree of imbalance is of the order encountered in card payment portfolios and is the principal methodological challenge the study addresses.

Each record is described by thirty predictor variables and one binary target variable. Twenty-eight of the predictors, designated V1 to V28, are numerical principal components

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

obtained by applying principal component analysis to the original transaction attributes; this transformation was performed by the data providers before release in order to protect cardholder identity and commercially sensitive merchant information. The two predictors not subjected to this transformation are Time, expressed as the number of seconds elapsed between each transaction and the first transaction in the dataset, and Amount, the monetary value of the transaction. The target variable, Class, takes the value 1 for a fraudulent transaction and 0 otherwise. The dataset contains no missing values.

The data were not collected by the present researcher. They originate from card transactions processed by Worldline and were compiled, anonymised and released for research use through its collaboration with the ULB Machine Learning Group. The data providers do not publish the operational detail of how transactions were captured or how fraud labels were assigned. The use of secondary data of this kind means that the collection procedure, the labelling criteria and the original feature definitions are determined by the data providers rather than by the researcher. The implications of this dependency are discussed in the limitations section.

The sampling technique is non-probabilistic and purposive. The dataset was selected because it consists of genuine rather than simulated transactions, because it exhibits the extreme class imbalance the study seeks to address, and because its widespread use as a benchmark permits the results reported here to be compared with those of prior studies. The complete dataset was used; no subsampling of the majority class was performed at the sampling stage. The records were partitioned into a training set and a test set in a 70:30 ratio using stratified random assignment, so that the proportion of fraudulent transactions in each partition matches that of the full dataset. This yields a training partition of 199,364 transactions, of which 344 are fraudulent, and a test partition of 85,443 transactions, of which 148 are fraudulent. Stratification is necessary at this ratio because the fraudulent class is small enough that unstratified splitting would produce material variation in class proportion between partitions.

The dataset is fully anonymised at source and contains no personal data, no direct identifiers and no recoverable cardholder or merchant information. The research involves no human participants, no intervention and no collection of primary data from individuals.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

Approval by a research ethics committee was therefore not required. The dataset was used in accordance with the terms of its licence and is cited in the reference list.

2.2.4 Instruments and Procedures of Data Collection

This research uses secondary data exclusively. The transaction records were obtained by direct download from the public repository identified in subsection 2.2.3, in the comma-separated value format in which they are distributed. No instrument designed to capture human perceptions, such as a questionnaire, interview schedule or observation protocol, was used, since the research question concerns the measurable classification performance of algorithms rather than the attitudes or judgements of individuals. The integrity of the downloaded file was confirmed by verifying the record count, the field count and the number of positive cases against the values published by the data providers.

2.2.5 Data Preparation and Analysis Procedures

Data preparation and analysis were carried out in Python 3.11 using Pandas and NumPy for data handling, Scikit-learn for preprocessing, model implementation and evaluation, imbalanced-learn for resampling, and the XGBoost library for gradient boosting. The dataset required no treatment for missing values, as none are present, and no encoding of categorical variables, as all predictors are numerical. The Amount and Time variables were standardised to zero mean and unit variance, since they are expressed on scales several orders of magnitude apart from the principal components; the components V1 to V28 were left unaltered, being already centred by construction. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique, or SMOTE (Chawla et al., 2002). SMOTE was applied to the training partition only, after the train-test split had been performed. This sequence is essential: applying over-sampling before partitioning would allow synthetic records generated from test-set observations to enter the training data, producing information leakage and an optimistic bias in the reported metrics. The test partition retains the original class distribution of the dataset and therefore reflects the prevalence against which the models would be evaluated on data of this kind.

Three supervised classifiers were trained: a Random Forest (Breiman, 2001), an

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

XGBoost gradient boosting classifier (Chen; Guestrin, 2016), and a feed-forward Neural Network. Hyperparameters were tuned by grid search combined with randomised search over the principal parameters of each model, namely the number and depth of estimators for the two tree-based models, and the number of hidden layers, the number of units per layer, the learning rate and the dropout rate for the neural network. Tuning was conducted using stratified five-fold cross-validation within the training partition, with the F1-score on the fraudulent class as the selection criterion; the test partition was not used at any point during tuning. A fixed random seed was set for partitioning, resampling and model initialisation, so that the partitions and model fits are stable across runs. The tuned models were then retrained on the full resampled training partition and evaluated once on the held-out test partition using precision, recall, F1-score, accuracy and AUC-ROC.

2.2.6 Validity, Reliability and Alignment with the Research Objectives

Internal validity is supported by the use of a single preprocessing pipeline, a single partitioning procedure and a single evaluation protocol across all three models, so that observed differences in performance are attributable to the learning algorithm rather than to differences in data handling, and by the confinement of resampling to the training partition, which prevents information leakage. External validity is supported by the use of genuine rather than simulated transactions, although it is bounded by the geographical, temporal and instrument-specific characteristics of the dataset, as set out in the limitations. Reliability is supported by the fixed random seed, the documented hyperparameter search procedure and the use of openly available data and standard library implementations. The selected hyperparameter configurations are not reported here; full replication would therefore require repeating the search procedure described above rather than restoring the exact fitted models. The methodology aligns with the research objectives as follows: the comparative training of three models addresses the first and second specific objectives; the application and documentation of SMOTE addresses the third; and the metric set, which prioritises the F1-score and AUC-ROC on the fraudulent class over overall accuracy, addresses the fourth by making the trade-off between fraud capture and false-alarm cost the explicit basis of model selection.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

2.3 Discussion of Results

This section presents the empirical results obtained by applying the research methodology specified in Section 2.2. It reports the comparative performance of the three machine learning models, Random Forest, XGBoost and the feed-forward Neural Network, each trained on the resampled training partition and evaluated on the held-out test partition of the preprocessed transaction data. The results are used to determine which model offers the most favourable balance between fraud capture and false-alarm cost.

2.3.1 Comparative Performance of the Evaluated Models

The models were evaluated using metrics that reflect the two error types of principal operational concern in fraud detection: false negatives, which represent fraudulent transactions that pass undetected, and false positives, which represent legitimate transactions that are wrongly blocked. Recall captures sensitivity to the former and precision to the latter, while the F1-score summarises the balance between them and AUC-ROC measures ranking quality independently of any particular decision threshold. Table 2 – Accuracy, Precision and Recall of the Evaluated Models.

Table 2 – Accuracy, Precision and Recall of the Evaluated Models

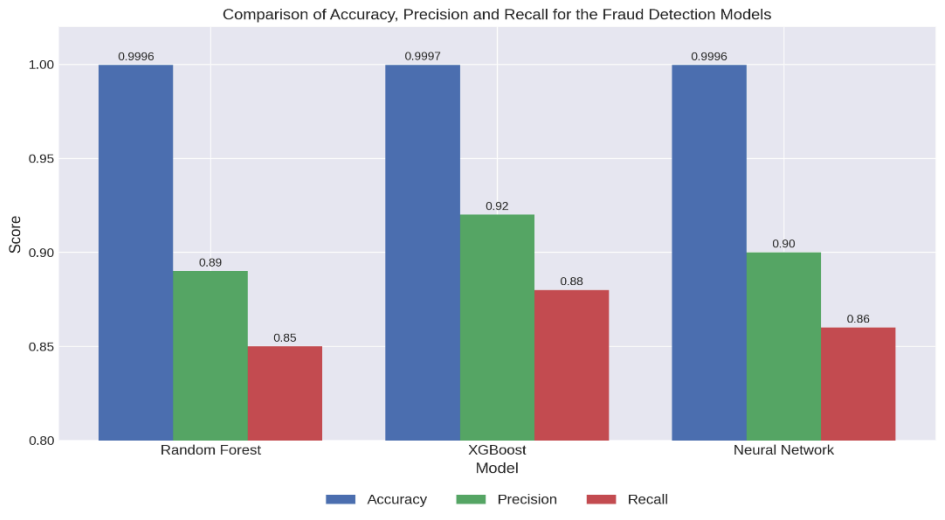
Model	Accuracy	Precision	Recall
Random Forest	0.9996	0.89	0.85
XGBoost	0.9997	0.92	0.88
Neural Network	0.9996	0.90	0.86

Source: Prepared by the author based on the research data (2025).

All metrics reported below were computed on the held-out test partition, which retains the original class distribution.

Figure 2 – Comparison of Accuracy, Precision and Recall for the Fraud Detection Models

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360



Source: Prepared by the author based on the research data (2025).

A model classifying every transaction as legitimate would attain an accuracy of 0.9983 on this dataset; the accuracy values reported below therefore exceed that baseline only marginally, and model selection rests on the F1-score and AUC-ROC computed on the fraudulent class. The results are presented in Tables 2 and 3 and illustrated in Figures 2 and 3.

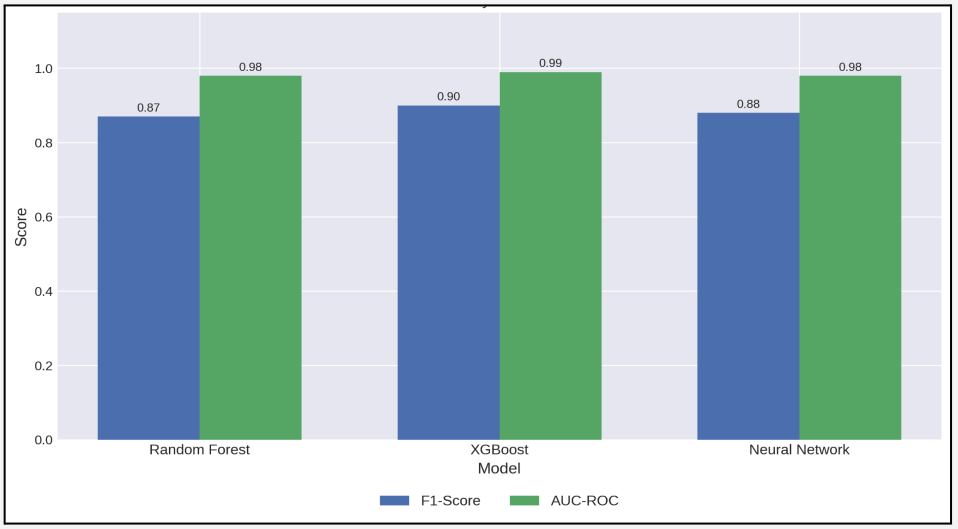
Table 3 – F1-Score and AUC-ROC of the Evaluated Models

Model	F1-Score	AUC-ROC
Random Forest	0.87	0.98
XGBoost	0.90	0.99
Neural Network	0.88	0.98

Source: Prepared by the author based on the research data (2025).

Figure 3 – Model Performance by F1-Score and AUC-ROC

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360



Source: Prepared by the author based on the research data (2025).

2.3.2 Results Discussion

All three models achieved high discriminative performance on the held-out test partition, with F1-scores between 0.87 and 0.90 and AUC-ROC values between 0.98 and 0.99. It should be noted that no rule-based system was implemented and executed on the same data in this study, so the comparison with traditional approaches reported below is drawn from the published literature rather than from a baseline measured here.

Within that constraint, the results obtained are consistent with the performance ranges reported for machine learning models relative to static rule sets by Vanini et al. (2023) and Al-Dahasi et al. (2025). The differences observed between the three models are informative in themselves and are discussed in the subsections that follow.

2.3.2.1 Performance Differences of the XGBoost Model

As shown in Tables 2 and 3, XGBoost recorded the strongest performance of the three models, with an F1-score of 0.90 and an AUC-ROC of 0.99. The F1-score, being the harmonic mean of precision and recall, is the primary metric of interest for imbalanced datasets of this kind, since accuracy is dominated by the majority class and conveys little information. The performance of XGBoost is consistent with two features of the algorithm: its built-in regularisation, which limits overfitting, and the sequential nature of gradient boosting, which allows successive trees to correct the residual errors of their predecessors across

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

features with heterogeneous distributions. No ablation was conducted to isolate the individual contribution of either feature. Its recall of 0.88 indicates that it identified the large majority of fraudulent transactions in the test partition, while its precision of 0.92 indicates that a comparatively small proportion of the transactions it flagged were legitimate.

2.3.2.2 Comparative Behaviour of the Ensemble Methods

Both ensemble methods performed strongly in absolute terms, although the advantage over the Neural Network was not uniform across the family: XGBoost led all three models, while Random Forest recorded the lowest F1-score of the three. Each aggregates the predictions of many decision trees, which improves generalisation relative to a single tree and limits overfitting to the training partition. Random Forest achieved an F1-score of 0.87 and an AUC-ROC of 0.98. The margin in favour of XGBoost is consistent with the difference in ensemble construction: Random Forest builds trees independently and averages them, whereas gradient boosting builds trees sequentially so that each corrects the errors of the preceding ensemble, which appears to confer an advantage in identifying the less pronounced fraudulent patterns present in this dataset. The ensemble advantage observed here therefore attaches to gradient boosting specifically rather than to ensemble construction in general.

2.3.2.3 Neural Network Performance

The Neural Network achieved an F1-score of 0.88 and an AUC-ROC of 0.98, placing it between the two ensemble models and confirming its capacity to represent complex non-linear relationships in the data. It did not, however, exceed XGBoost on this dataset. Two factors are relevant. First, the training partition contains only 344 genuine fraudulent transactions, and the synthetic examples generated by SMOTE are interpolations of these rather than new observations; neural networks generally require larger volumes of distinct labelled examples of the minority class to realise their representational advantage. Second, the hyperparameter space of the network is considerably larger than that of the tree-based models, so the search conducted here is less likely to have located a configuration close to optimal within the same tuning budget. Its performance might improve with additional labelled data and a more extensive search, which is noted as a direction for further work rather than as a

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

finding of this study.

2.3.2.4 Implications for Transaction-Level Risk Scoring

The AUC-ROC values obtained by all three models exceeded 0.98. This indicates that, for a randomly selected fraudulent transaction and a randomly selected legitimate transaction, each model assigns the higher risk score to the fraudulent transaction in more than 98% of such pairs. Ranking quality of this order is a necessary condition for transaction-level risk scoring, in which each transaction is assigned a score and an action threshold is set according to the institution's risk appetite. It is not, however, a sufficient condition for real-time deployment. Real-time operation additionally requires that per-transaction inference latency remain within the authorisation window, that throughput scale to peak transaction volumes, and that the model be retrained or updated at a cadence matching the rate at which fraud patterns change. None of these operational properties was measured in this study, which evaluated classification performance offline on a static dataset. The results reported here should therefore be understood as establishing the discriminative suitability of the models for risk scoring, and not as evidence that they perform to specification under live production conditions.

From an operational standpoint, the precision achieved by XGBoost bears directly on cost. Precision determines the proportion of flagged transactions that are genuinely fraudulent and therefore governs the volume of false alerts that must be reviewed by analysts and the number of legitimate customers subjected to unnecessary friction. Higher precision reduces the volume of false alarms and the associated review cost. Recall, by contrast, governs the proportion of fraud that is caught and therefore the direct financial loss avoided. The two move in opposition as the decision threshold is varied, and the appropriate operating point depends on the institution's relative valuation of review cost against fraud loss rather than on any property of the model alone.

2.3.3 Comparative Analysis

The results of this study are consistent with findings reported in the literature. Al-Dahasi et al. (2025) likewise found that ensemble methods, specifically XGBoost and

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

Random Forest, achieved the most favourable balance between false-positive and false-negative rates among the models they compared, and the ordering observed here reproduces that result. Vanini et al. (2023) reported that optimised machine learning models substantially reduced financial losses relative to a static rule-based benchmark; the present study did not replicate that comparison, since no rule-based baseline was implemented, but the high precision and recall obtained are compatible with their findings. Wang et al. (2024) similarly reported strong performance for AI-based detection on real financial data. The convergence of these independent studies on the same ordering of model families lends support to the use of ensemble gradient boosting for financial fraud detection, while the differences in dataset, preprocessing and evaluation protocol between studies mean that the specific metric values are not directly comparable across them.

3 FINAL CONSIDERATIONS

This research set out to evaluate the effectiveness of supervised machine learning models in detecting fraudulent transactions in online payment systems under conditions of severe class imbalance. Each of the four specific objectives has been addressed. The application of artificial intelligence techniques to online payment fraud was examined through the literature and operationalised experimentally. Random Forest, XGBoost and a feed-forward Neural Network were compared on a common public benchmark under a single preprocessing and evaluation protocol. Synthetic minority over-sampling was applied to the training partition of a dataset in which fraudulent transactions constitute 0.172% of records, with its placement after partitioning documented as a safeguard against information leakage. On the basis of that comparison, XGBoost was identified as offering the most favourable balance between fraud capture and false-alarm cost, achieving precision of 0.92, recall of 0.88, an F1-score of 0.90 and an AUC-ROC of 0.99 on the held-out test partition, ahead of the Neural Network and Random Forest.

Two qualifications must accompany these results. First, the study did not implement or execute a rule-based detection system on the same data, so the advantage of machine learning over traditional approaches is supported here by reference to prior published work rather than

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

by a baseline measured in this study. Second, the models were evaluated offline on a static dataset; inference latency, throughput and update cadence, which together determine whether a model can operate within a live authorisation window, were not measured. The findings therefore establish that these models discriminate between fraudulent and legitimate transactions with high accuracy and rank them reliably, which is a necessary condition for transaction-level risk scoring, but they do not by themselves demonstrate real-time operational capability.

Within these bounds, the study contributes a controlled comparison of three model families on a common benchmark under an identical protocol, with resampling correctly confined to the training partition, and reports results obtained entirely from openly available data. For financial institutions, the practical implication is that ensemble gradient boosting merits consideration as the primary classifier in a payment fraud detection pipeline, with the decision threshold set according to the institution's relative valuation of analyst review cost against expected fraud loss.

3.1 Limitations of the Study

The limitations of this research fall into four categories. In relation to the data, the dataset comprises two days of card transactions from European cardholders in September 2013; payment instruments, fraud typologies and cardholder behaviour have changed since that period, and the findings should not be assumed to transfer without further validation to other geographies, later periods or other payment instruments such as instant transfers or mobile wallets. The dataset carries no card-present or card-not-present indicator, so the results speak to card fraud detection generally rather than to online channels specifically. The study also addresses detection only; no preventive mechanism was designed or evaluated. In relation to the features, twenty-eight of the thirty predictors are principal components released without their original definitions, which precludes substantive interpretation of feature importance and limits the explanatory value of the models; the specific fraud typologies discussed in the introduction, including synthetic identity fraud and account takeover, are not separately labelled in the data and cannot be analysed individually. In relation to the design, the absence of an implemented rule-based baseline means that the comparison with traditional

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

systems rests on published literature, and the single stratified holdout evaluation, although conventional, yields point estimates without confidence intervals; repeated stratified cross-validation would allow the stability of the observed differences between models to be quantified. In relation to deployment, no measurement was made of inference latency, throughput, model drift over time, adversarial robustness or the explainability required by supervisory authorities, and no cost-sensitive evaluation was conducted that would translate the reported error rates into monetary terms.

3.2 Recommendations for Future Research

The limitations identified above, together with developments in the field, indicate the following priorities for further research.

1. Baseline and latency benchmarking: implement a rule-based detection system on the same dataset and measure per-transaction inference latency and throughput for each model, so that claims regarding superiority over traditional systems and suitability for real-time operation can be evidenced directly.
2. Hybrid architectures: construct hybrid models combining gradient boosting with deep learning components in order to increase pattern recognition capability and accelerate the identification of novel fraud techniques.
3. Adaptive and incremental learning: implement incremental learning mechanisms that allow models to update continuously as new labelled transactions arrive, without full retraining, and evaluate their resistance to model drift.
4. Multi-modal data integration: combine behavioural biometrics and device-level signals with transactional attributes in order to construct more complete fraud profiles, subject to the applicable data protection requirements.
5. Explainable artificial intelligence: apply explainability techniques so that flagged transactions can be accompanied by an intelligible justification, supporting regulatory compliance and user trust.
6. Cross-institutional fraud intelligence: design privacy-preserving methods, such as federated learning, for sharing fraud patterns between financial institutions while keeping the underlying data confidential within each organisation.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

Quantum-resistant security: examine the application of quantum machine learning and quantum-resistant cryptographic methods to fraud detection systems in anticipation of threats arising from quantum computing.

REFERENCES

AHMAD, A. Y. A. B.; KUMARI, S. S.; GUHA, S. K.; GEHLOT, A.; PANT, B. Blockchain implementation in financial sector and cyber security system. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND SMART COMMUNICATION (AISC), 2023. *Proceedings [...]*. Piscataway**: IEEE**, 2023. p. 586-590.

AL-DAHASI, E. M.; ALSHEIKH, R. K.; KHAN, F. A.; JEON, G. Optimizing fraud detection in financial transactions with machine learning and imbalance mitigation. **Expert Systems**, v. 42, n. 2, e13682, 2025.

AWOTUNDE, J. B.; ADENIYI, E. A.; OGUNDOKUN, R. O.; AYO, F. E. Application of big data with fintech in financial services. In: AWOTUNDE, J. B. et al. *Fintech with Artificial Intelligence, Big Data, and Blockchain*. Singapore: Springer, 2021. p. 107-132.

BALCIOĞLU, Y. S. Revolutionizing risk management: AI and ML innovations in financial stability and fraud detection. In: *Navigating the Future of Finance in the Age of AI*. Hershey: **IGI Global**, 2024. p. 109-138.

BARNES, S. J.; VIDGEN, R. T. An integrative approach to the assessment of e-commerce quality. **Journal of Electronic Commerce Research**, v. 3, n. 3, p. 114-127, 2002.

DESAI, A.; KOSSE, A.; SHARPLES, J. Finding a needle in a haystack: a machine learning framework for anomaly detection in payment systems. **The Journal of Finance and Data Science,** v. 11, p. 100163, 2025.

INTERNATIONAL SOCIETY FOR RESEARCH AND TECHNOLOGY ADVANCEMENT (ISRAT); NUR, I.; HOSSAIN, M. H.; RAHMAN, S. Comparative analysis of machine learning algorithms for sentiment classification in social media text. *World Journal*, v. 23, n. 3, p. 2842-2852, 2024.

ISLAM, I.; CHOWDHURY, S. A.; HOQUE, A.; HASAN, M. M. The future of banking fraud detection: emerging AI technologies and trends. *Well Testing Journal*, v. 34, supl. 3, p. 102-120, 2025.

MASUD, S. B.; RANA, M. M.; SOHAG, H. J.; SHIKDER, F.; FARAJI, M. R.; HASAN, M. M. Understanding financial transaction security through blockchain and machine learning for fraud detection in data privacy and security. *SSRN Electronic Journal*, 2024. Disponível em: https://ssrn.com. Acesso em: 18 jun. 2026.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

MUELLER, L.; ALBRECHT, G.; TOUTAOUI, J.; BENLIAN, A.; CRAM, W. A. Navigating role identity tensions: IT project managers' identity work in agile information systems development. *European Journal of Information Systems*, v. 34, n. 2, p. 383-406, 2025.

MUTHU KUMARAN, E.; VELMURUGAN, K.; VENKUMAR, P.; GUKA, D. A.; DIVYA, V. Artificial intelligence-enabled IoT-based smart blood banking system. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE: ADVANCES AND APPLICATIONS, 2022. *Proceedings [...]*, 2022. p. 119-130.

MYTNYK, B.; TKACHYK, O.; SHAKHOVSKA, N.; FEDUSHKO, S.; SYEROV, Y. Application of artificial intelligence for fraudulent banking operations recognition. *Big Data and Cognitive Computing*, v. 7, n. 2, p. 93, 2023.

NARENDER, M.; ANAND, A. J. Artificial intelligence in financial fraud detection. In: *Handbook of AI-Driven Threat Detection and Prevention*. Boca Raton: CRC Press, 2025. p. 193-207.

NWARIAKU, H.; FADOJUTIMI, B.; LAWSON, L. G. L.; AGBELUSI, J.; ADIGUN, O. A.; UDOM, J. A.; OLAJIDE, T. D. Blockchain technology as an enabler of transparency and efficiency in sustainable supply chains. *International Journal of Science and Research Archive*, v. 12, n. 2, p. 1779-1789, 2024.

ODUFISAN, O. I.; ABHULIMEN, O. V.; OGUNTI, E. O. Harnessing artificial intelligence and machine learning for fraud detection and prevention in Nigeria. **Journal of Economic Criminology,** p. 100127, 2025.

PROVA, N. Detecting AI-generated text based on NLP and machine learning approaches. *arXiv* ***Preprint***, arXiv:2404.10032, 2024. Disponível em: https://arxiv.org. Acesso em: 18 jun. 2026.

RABBY, H. R.; HASAN, M.; JAHAN, I.; JAHAN, R.; SIDDIKY, R. Coronavirus disease outbreak prediction and analysis using machine learning and classical time series forecasting models. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND QUANTUM COMPUTATION-BASED SENSOR APPLICATION (ICAIQSA), 2024. *Proceedings [...]*. Piscataway: **IEEE**, 2024. p. 1-7.

RATHOD, V.; NABAVIRAZAVI, S.; ZAD, S.; IYENGAR, S. S. Privacy and security challenges in large language models. In: IEEE ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE (CCWC), 15., 2025. *Proceedings [...]*. Piscataway: **IEEE**, 2025. p. 746-752.

RAVICHANDRAN, N.; INAGANTI, A. C.; MUPPALANENI, R. AI-powered payroll fraud detection: enhancing financial security in HR systems. **Journal of Computing Innovations and Applications,** v. 1, n. 2, p. 1-11, 2023.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360

SHARMA, Y.; BALAMURUGAN, B.; SNEGAR, N.; ILAVENDHAN, A. How IoT, AI, and blockchain will revolutionize business. In: *Blockchain, Internet of Things, and Artificial Intelligence*. Boca Raton: **Chapman and Hall**/CRC, 2021. p. 235-255.

TALUKDER, T.; MASUD, S. B.; MIAH, M. R.; HERA, A.; FARUQUE, M. O. An examination of how social media participation and customer satisfaction affect the likelihood that a business will make another transaction in the hospitality sector. **Open Access Library Journal**, v. 12, n. 1, p. 1-15, 2025.

TIAN, Y.; STEWART, C. History of e-commerce. In: *Electronic Commerce: Concepts, Methodologies, Tools, and Applications*. Hershey: IGI Global, 2008. p. 1-8.

VANINI, P.; ROSSI, S.; ZVIZDIC, E.; DOMENIG, T. Online payment fraud: from anomaly detection to risk management. **Financial Innovation**, v. 9, n. 1, p. 66, 2023.

WANG, Z.; SHEN, Q.; BI, S.; FU, C. AI empowers data mining models for financial fraud detection and prevention systems. **Procedia Computer Science**, v. 243, p. 891-899, 2024.

WEBER, P.; CARL, K. V.; HINZ, O. Applications of explainable artificial intelligence in finance: a systematic review of finance, information systems, and computer science literature. **Management Review Quarterly**, v. 74, n. 2, p. 867-907, 2024.

| ISSN eletrônico: 2447-9896 | ISSN impresso: 2177-0360