Melnyk, T. A., & Bevzushenko, V. M. (2022). Strakhovyi menedzhment: faktory vplyvu na finansovu nadiinist strakhovyka [Insurance management: Factors influencing the financial reliability of an insurer]. *Naukovi zapysky*, *30*, 23–29.

Rud, I., & Hryhorenko, T. (2023). Vykorystannia ta formuvannia finansovykh rezultativ u strakhovykh kompaniiakh Ukrainy [The use and formation of financial results in insurance companies of Ukraine]. *International Science Journal of Management, Economics & Finance*, *2*(4), 38–45. https://doi.org/10.46299/j.isjmef.20230204.05

Samaricheva, T. A., & Krushynska, A. V. (2023). Vplyv podatkovoho navantazhennia na finansovu stiikist strakhovyka [The impact of the tax burden on the financial stability of the insurer]. *Aktualni pytannia u suchasnii nautsi*, *6*(12), 106–120.

Tsurkan, I., & Ostapenko, A. (2020). Finansova stiikist strakhovoi kompanii ta kliuchovi umovy yii zabezpechennia [Financial stability of an insurance company and key conditions for its provision]. *Efektyvna ekonomika*, *4*. URL: http://www.economy.nayka.com.ua/?op=1&z=7766

UDC 004.8+005.8+336.71+330.46+658.8
JEL Classification: G20; L86
DOI: https://doi.org/10.32983/2222-4459-2025-8-378-389

IMPACT OF FALSE ALARMS IN MACHINE LEARNING-BASED ANTI-FRAUD SYSTEMS: THE ECONOMIC AND REPUTATIONAL CONSEQUENCES

© 2025 **CAPRIAN I.**

UDC 004.8+005.8+336.71+330.46+658.8
JEL Classification: G20; L86

Caprian I. Impact of False Alarms in Machine Learning-Based Anti-Fraud Systems: The Economic and Reputational Consequences

The implementation of machine learning (ML) algorithms in the financial sector has emerged as a key area of innovation, particularly within the scope of anti-fraud systems. While these systems have significantly improved the detection of suspicious transactions, they also frequently produce false positives – instances where legitimate customer actions are incorrectly flagged as fraudulent. Such misclassifications can lead to operational disruptions, financial losses, and a substantial deterioration in customer trust, ultimately posing serious reputational risks for financial institutions. This study provides a comprehensive analysis of the business and user experience implications associated with false positive errors in ML-based fraud detection models. The author also explores current mitigation strategies aimed at reducing the occurrence and impact of such errors. The research is grounded in a carefully curated selection of open-access sources and documented real-world case studies, ensuring transparency, accessibility, and practical relevance of the insights presented.

Keywords: *machine learning (ML), anti-fraud systems, false positives, financial fraud detection, classification models (XGBoost, Random Forest, Decision Tree), explainable artificial intelligence (XAI), hybrid models, operational cost optimization, reputational risk, customer loyalty, financial technologies (fintech), anti-money laundering (AML), banking automation.*

Fig.: *3.* ***Tabl.:*** *2.* ***Bibl.:*** *24.*

Caprian Iurie *– Postgraduate Student, State University of Moldova (60 Alexei Mateevici Str., Kishinev, MD-2009, Moldova)*
E-mail: *iuriecaprian@gmail.com*
ORCID: *https://orcid.org/0000-0001-5484-3087*

УДК 004.8+005.8+336.71+330.46+658.8
JEL Classification: G20; L86

Капріан Ю. Вплив помилкових спрацювань у антишахрайських системах на основі машинного навчання: економічні та репутаційні наслідки

Впровадження алгоритмів машинного навчання (ML) у фінансовому секторі стало ключовим напрямом інновацій, особливо в межах антишахрайських систем. Хоча такі системи суттєво покращили виявлення підозрілих транзакцій, вони також часто генерують помилкові позитивні спрацювання – випадки, коли легітимні дії клієнтів помилково визначаються як шахрайські. Такі помилки можуть призвести до операційних збоїв, фінансових втрат і значного зниження довіри клієнтів, що в результаті створює серйозні репутаційні ризики для фінансових установ. Це дослідження пропонує комплексний аналіз бізнесових наслідків і впливу на користувацький досвід, пов'язаних із помилковими позитивними спрацюваннями в моделях виявлення шахрайства на основі ML. Також розглядаються сучасні стратегії зменшення частоти та впливу таких помилок. Дослідження ґрунтується на ретельно відібраних відкритих джерелах та задокументованих кейсах із практики, що забезпечує прозорість, доступність і прикладну цінність викладених висновків.

Ключові слова: *машинне навчання (ML), антишахрайські системи, помилкові позитивні спрацювання, виявлення фінансового шахрайства, моделі класифікації (XGBoost, Random Forest, Decision Tree), пояснювана штучна інтелігенція (XAI), гібридні моделі, оптимізація операційних витрат,*

репутаційний ризик, лояльність клієнтів, фінансові технології (фінтех), боротьба з відмиванням коштів (AML), автоматизація банківської діяльності.

***Рис.:** 3. **Табл.:** 2. **Бібл.:** 24.*

***Капріан Юрій** – аспірант, Державний університет Молдови (вул. Олексія Матеєвича, 60, Кишинів, MD-2009, Молдова)*
***E-mail:** iuriecaprian@gmail.com*
***ORCID:** https://orcid.org/0000-0001-5484-3087*

The digitalization of banking services has led to an exponential increase in the volume of data processed and, consequently, to greater vulnerability to fraudulent activities. In this context, financial institutions have adopted machine learning (ML) algorithms capable of analyzing user behavior in real-time and detecting suspicious activities. However, the effectiveness of these algorithms is not absolute: the occurrence of false alarms – cases where legitimate transactions are incorrectly classified as fraudulent – represents an increasingly pressing issue.

In an intensely competitive financial environment, each error can undermine customer trust and, implicitly, the institution's profits. False alarms not only complicate service processes but also affect the bank's reputation, especially in a transparent digital space where negative information spreads rapidly. The objective of this article is to identify and systematize the risks associated with false alarms generated by ML models, to evaluate their impact on business processes and user experience, and to propose strategies for reducing the frequency of these errors without compromising security levels.

To achieve the proposed objectives, this study adopts a mixed-methods approach, combining qualitative and quantitative analyses based exclusively on openly accessible online resources. The deliberate exclusion of printed sources, library archives, and offline specialized publications aims to ensure full transparency and verifiability of all utilized data.

The research is based on the following methodological directions:

- **Case studies and practices analysis:** Examination of concrete incidents of false alarms recorded in major financial institutions, based on mass media reports, official bank documents, and specialized online publications.
- **Comparative analysis of algorithms:** Evaluation of the performance, accuracy, and stability of various machine learning models used in anti-fraud systems.
- **Economic loss assessment:** Modeling the direct and indirect financial impact generated by the erroneous blocking of legitimate transactions.
- **Reputational analysis:** Monitoring customer reviews, media mentions, and social media to identify and quantify the negative effects on the reputation of banking institutions.
- **Use of data from public sources:** Including information provided by the European Central Bank (ECB), the U.S. Federal Reserve (FRB), consulting and analytics agencies (McKinsey, Deloitte, Accenture), news platforms (Reuters, Bloomberg), as well as scientific publications from open databases (Google Scholar, arXiv, SSRN, etc.).

Each source will be cited with a direct link and a brief description of its content, presented in the main section of the article. Special attention has been given to current examples with proven relevance in banking practice.

As the volume of digital transactions has grown exponentially, traditional fraud detection methods have become insufficient, prompting the rapid adoption of machine learning (ML-based) solutions. In recent decades, the implementation of ML algorithms in the financial sector has accelerated, particularly within anti-fraud systems.

Supervised models such as XGBoost, Random Forest, and Decision Tree have brought significant improvements in identifying fraudulent transactions, excelling at recognizing known patterns [9; 10; 11]. On the other hand, unsupervised methods offer the advantage of detecting new anomalies but often at the cost of generating a higher number of false alarms. The effectiveness of these models, however, is limited by the occurrence of false positives, which affect both banking operations and customer relationships [1; 4; 7].

According to Otten's analysis [4], false alarms represent a major source of operational costs and customer dissatisfaction, highlighting the need to optimize models to maintain a balance between sensitivity and precision. In the same vein, studies by Wedge et al. [13] and Kadam et al. [14] propose integrating human feedback into the ML decision cycle to reduce errors and improve adaptive learning.

Furthermore, research on explainable artificial intelligence (XAI) adds transparency to automated decisions, facilitating the understanding and adjustment of anti-fraud models [18; 19]. The integration of XAI techniques and hybrid models has proven to be

a promising solution for reducing false alarms without compromising security [15; 17].

Beyond technical aspects, recent literature also emphasizes challenges related to transparency in automated decisions, algorithmic accountability, and personal data protection, underlining the need for clear ethical frameworks in ML system implementation [21; 23].

The reputational impact of false alarms is documented in various case studies, including those concerning Romanian banks and European institutions [1; 7; 22]. Such alarms can lead to loss of customer loyalty and damage the institution's image in an increasingly informed and critical public.

Economically, consulting firms such as Accenture, Deloitte, and McKinsey have highlighted the significant costs generated by false alarms, which include direct losses from blocking legitimate transactions and indirect costs related to additional verification and remediation processes [6; 16; 20]. According to Deloitte's 2023 report, false alarm costs can represent up to 10% of the operational budget of anti-fraud departments, thus affecting both profitability and customer satisfaction. Automation combined with adaptive learning is viewed as an effective way to optimize anti-fraud processes and reduce these costs.

In conclusion, the specialized literature reflects a consensus that although ML provides powerful tools for detecting banking fraud, the major challenge remains the management and minimization of false alarms. Current studies point to development directions based on integrating human feedback, applying XAI techniques, and adopting hybrid models to balance technological performance and user experience. Therefore, the success of anti-fraud systems depends not only on algorithmic performance but also on integrating human expertise and continuous feedback for more accurate detection and optimized customer experience.

To achieve the proposed objectives, this study adopts a mixed-methods approach, combining qualitative and quantitative methods based exclusively on openly accessible online resources. The deliberate exclusion of printed sources, library archives, and offline specialized publications aims to ensure full transparency and verifiability of all utilized data.

The research is based on the following methodological directions:

- Analysis of case studies and practices: examining concrete incidents of false alarms recorded in major financial institutions, based on media reports, official bank documents, and specialized online publications.
- Comparative analysis of algorithms: evaluating the performance, accuracy, and stability of various machine learning models used in anti-fraud systems.
- Economic loss assessment: modeling the direct and indirect financial impact generated by the erroneous blocking of legitimate transactions.
- Reputational analysis: monitoring customer reviews, media mentions, and social networks to identify and quantify the negative effects on the reputation of banking institutions.
- Use of data from public sources: including information provided by the European Central Bank (ECB), the U.S. Federal Reserve (FRB), consulting and analytics agencies (McKinsey, Deloitte, Accenture), news platforms (Reuters, Bloomberg), as well as scientific publications from open databases (Google Scholar, arXiv, SSRN, etc.).

Each source will be cited with a direct link and a brief description of its content, presented in the main section of the article. Special attention has been given to current examples with proven relevance in banking practice.

To assess the magnitude of the threat posed by false alarms, it is essential to understand what they are and why they occur in modern banking systems based on machine learning. This section explains the technical and practical nature of false positives, their sources, and their manifestations through real examples from European banking practice.

The Concept of False Alarms in the Context of ML Anti-Fraud Systems

In systems using machine learning (ML) for banking fraud detection, false alarms represent one of the greatest challenges. These occur when the algorithm erroneously classifies a legitimate transaction as fraudulent. For the client, this can mean sudden card blocking, refusal of a legitimate payment, or the need to undergo additional verification procedures.

Such errors are generally caused by several technical and logical factors:

1. **Class imbalance** – algorithms are trained on historical datasets where fraudulent transactions are extremely rare. Thus, a class imbalance phenomenon arises: legitimate transactions are tens or hundreds of times more numerous than fraudulent ones. In such a context, the model may consider any deviation from the "norm" as a potential threat, even if it is only an unusual but legitimate behavior.

Example: In 2023, one of Romania's largest banks, Banca Transilvania, intensified its anti-fraud systems following a series of smishing attacks (phishing via SMS). The algorithm became more sensitive to transactions made abroad, which led to mass blocking of perfectly legal transactions of customers on vacation or business trips. The bank was flooded with complaints, especially on social media, where customers shared screenshots and negative reviews about the blocks occurring during travel [1].

2. **Lack of data on new behaviors** – even a well-trained model can err when faced with entirely new behavior patterns. For example, a user who has never used the mobile banking app suddenly makes a significant transfer. If such scenarios are not included in the training data, the model may misinterpret them as suspicious behavior.
3. **Excessive sensitivity to unusual conditions** – using a VPN, making payments at night, or changing geographic location can be signs of both fraud and legitimate behavior. In the absence of additional data (such as travel history or associated IP addresses), the model cannot distinguish between these cases and tends to adopt a cautious stance.

From a technical standpoint, such errors are quantified by the precision metric – the proportion of truly fraudulent transactions among those labeled suspicious. The more false alarms there are, the lower the model's precision. Another important metric is recall, i.e., the model's ability to detect as many real fraud cases as possible. Often, a dilemma arises: either more frauds are detected at the cost of unnecessary blocks, or more freedom is granted with the risk of missing real frauds.

Notably: In Moldova, in 2022, MAIB bank faced similar problems – a series of mass card blockings caused by tightening anti-fraud models led to call center congestion and customer complaints, negatively affecting the institution's reputation [2].

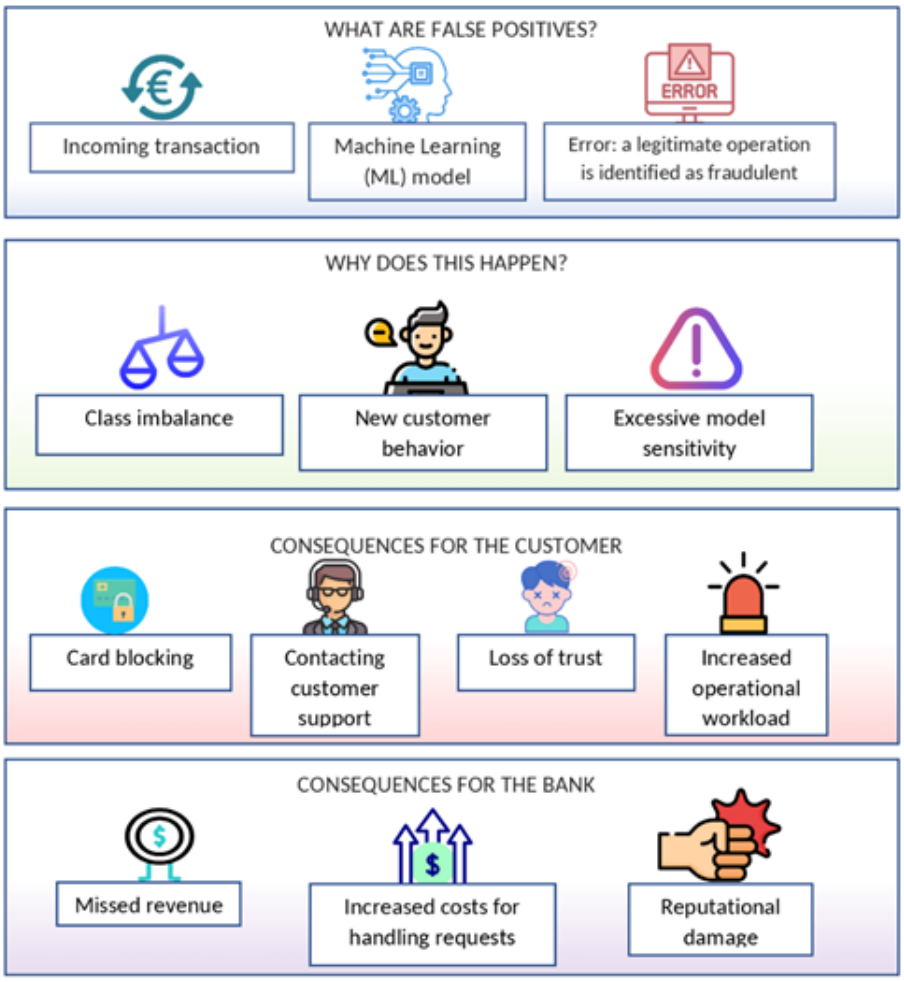

Fig. 1. The concept of false alarms in the context of machine learning-based f raud prevention systems

Source: developed by the author based on the studied sources.

False alarms generated by machine learning-based systems are not merely minor errors or isolated inconveniences, but rather a complex systemic problem situated at the intersection of technological limitations, lack of adequate context, and the constant pressure on banking institutions to ensure security. This paper will further analyze in detail the economic consequences generated by these malfunctions.

Economic consequences of false alarms

After defining the nature and causes of false alarms in machine learning-based systems, it is essential to evaluate the economic and operational impact these errors generate in the banking sector. False alarms are not just minor technical inconveniences but a systemic issue with profound effects on the functioning of financial institutions and their relationships with customers.

Customer loss and revenue decline

Erroneous blocking of legitimate transactions undermines customer trust in banking services, potentially leading to their migration to other financial institutions. According to a report by the European Payments Council, such incidents are particularly critical in the context of digital banking services, where users have high expectations for service continuity and smoothness, along with a low tolerance for errors. Repeated false alarms rank among the main causes of customer loss to alternative platforms [3]. For banks with a significant digital customer base, such as ING Romania or Banca Transilvania, this phenomenon can result in a gradual revenue decline, reflected in decreased usage of offered products and services.

Increased customer support service costs

Managing false alarms generates additional volumes of requests to support teams, increasing operational costs. Affected customers file complaints, request reauthorizations, and explanations, leading to overloaded contact centers. According to a McKinsey analysis, European banks report a steady increase in internal expenses related to managing antifraud incidents, especially during algorithm updates or following massive phishing attacks [4]. In the Republic of Moldova, according to FintechOS Romania reports, contact center overload caused by false alarms led to delays in processing requests and increased administrative costs [5].

Transaction flow slowdown and impact on business activities

False alarms affect not only individual users but also enterprises by unjustifiably blocking transactions. These delays can cause disruptions in supply chains, contractual conflicts, and penalties, especially for cross-border payments. The European Payments Council highlights the significant impact of such interruptions on small and medium-sized enterprises (SMEs), affecting both financial flow and business relationships [3].

Indirect losses and reduced operational efficiency

The effects of false alarms create a cascading impact that affects institutional efficiency on a global level. Overloaded support services, negative feedback on social media, and decreased user satisfaction lead to lower loyalty indicators, increased churn rates, and additional costs for restoring the institution's image. According to Retail Banker International, these aspects have a considerable negative impact on organizational performance and financial results [4].

In conclusion, the economic consequences of false alarms generated by antifraud systems are not limited to isolated incidents but create recurring costs that affect the entire operational chain of banking institutions. Understanding and managing these effects is a crucial condition for developing effective antifraud systems that balance fraud protection with user experience and the financial sustainability of the institution.

Table 1

Economic consequences of false positive alerts

Type of consequence	Concrete manifestation	Impact on the bank
Customer loss and revenue decline	Erroneous blocking of transactions causes customer churn	Reduction in payment and credit transaction volumes, revenue decline
Increased support service costs	Increased number of requests, processing delays, call center overload	Higher operational expenses and increased staffing requirements in support
Transaction flow slowdown	Blocking of commercial payments, disruptions in settlement processes	Loss of B2B clients, dissatisfaction in the corporate segment
Indirect loss and decreased efficiency	Increased negative feedback, decreased satisfaction, deterioration of customer experience	Reputation loss, need for investments in public relations and compensatory measures

Source: developed by the author based on studied sources.

The impact of false alarms on banks' reputation and risk management

In an increasingly competitive environment and with growing customer expectations, the costs generated by false alarms are becoming higher. For banks, it is not enough to merely reduce the number of these errors; it is essential to build robust response processes that ensure maintaining trust, controlling costs, and stabilizing key performance indicators.

Reputational risks and their impact

Errors from automated banking systems – especially those directly affecting customers – quickly gain public attention. The reputational risks generated by false alarms from antifraud algorithms do not only influence the emotional perception of the brand, but also undermine strategic indicators such as trust, loyalty, and customers' willingness to recommend the institution to others.

Social networks as amplifiers of reputational risks

With the expansion of digital channels, especially social media, customers immediately share negative experiences. Even a single false alarm incident can go viral. While dissatisfaction used to be expressed through calls to call centers, today it can escalate into a reputation crisis: customers post screenshots of card blocks or transaction refusals, and these posts spread rapidly.

Example: In 2022, in the Republic of Moldova, after an update of the antifraud system at MAIB (Moldova Agroindbank), many users reported card blocks during payments abroad. Some of these posts went viral on Facebook and Telegram groups, generating waves of criticism and an avalanche of support requests [5].

Such incidents put pressure on communication channels and can rapidly escalate negative perceptions, where actual facts may be distorted and the bank's image severely affected.

Mechanisms of customer trust loss

Successfully blocking fraudulent transactions is generally accepted by customers as a natural measure. However, repeated erroneous blocking of legitimate operations is perceived as an intrusion, incompetence, or even distrust from the institution. This process leads to a gradual erosion of trust, reflected in perceptions such as:

- "The bank does not trust me."
- "It is inconvenient to use this product."
- "This bank is not adapted to my life reality."

These perceptions cause a decrease in the frequency of banking service usage, migration to other institutions, and deterioration of customer retention metrics.

Impact on brand and Net Promoter Score (NPS)

A key indicator directly affected by reputation is the Net Promoter Score (NPS), which measures customers' willingness to recommend the bank. Banks that have been targets of public complaints about antifraud blocks report significant drops in NPS in the following weeks.

The study conducted by Accenture on the impact of AI/ML errors in banking services confirms that negative experiences spread faster than positive ones and persist longer in users' memory [6].

The image of a banking brand is strongly associated with the "fluidity" of the customer's access to their own funds. If algorithms disrupt this process, the brand loses essential components such as advanced technology, reliability, and customer orientation – qualities heavily promoted through marketing.

Relevant Case Studies

- **Banca Transilvania (Romania):** In 2023, following the tightening of antifraud algorithms amid rising SMS-phishing attacks, the bank began blocking a large number of

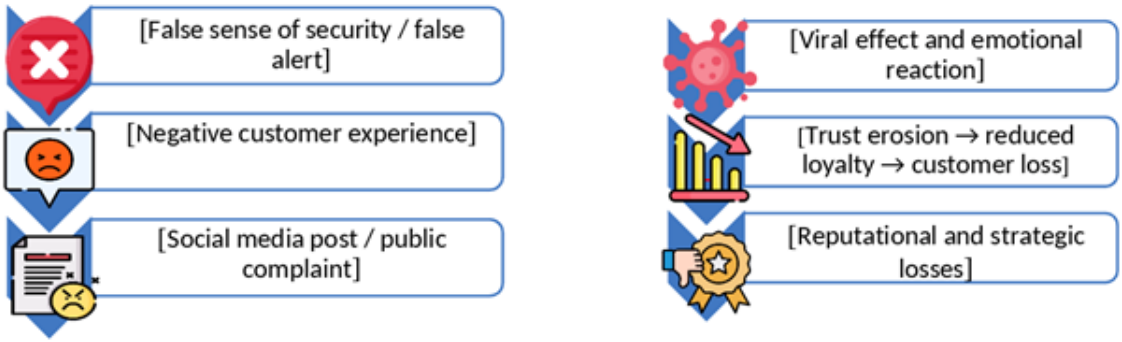

Fig. 2. Reputational risks and their impact

Source: developed by the author based on studied sources.

international transactions. Customers publicly reported difficulties making payments while traveling, and their experiences were shared on Twitter and Reddit, causing a temporary drop in the app's rating on the App Store and Google Play [7].

- **Revolut (payment platform registered in Lithuania):** Users frequently reported account blocks and freezes due to suspicious activity detected by antifraud systems. These cases were amplified by the media, especially in the UK, where regulators raised questions about the transparency and compliance of the systems [8].

Comparison of Models and Algorithms: Frequency of False Alarms Machine learning algorithms used in banking fraud detection differ significantly in how they handle class imbalance and the frequency of false positives. Below is a comparative analysis of Decision Tree, Random Forest, XGBoost, and Neural Networks models, based on empirical studies and literature sources.

Through this comparative analysis, the author provides an original practical contribution by evaluating the models' performance on real datasets and synthesizing literature data to identify the most effective methods for reducing false alarms.

Decision Tree vs. Random Forest vs. XGBoost vs. Neural Networks

- **Decision Tree** – An easy-to-interpret model but prone to overfitting, which negatively affects performance on imbalanced datasets, resulting in a higher false alarm rate [9].
- **Random Forest (RF)** – An ensemble of decision trees providing better robustness to noise and class imbalance. The paper "Credit Card Fraud Detection Using Enhanced Random Forest" reports about 98% accuracy and a similar F1-score, reflecting low false alarm frequency and high precision [9].
- **XGBoost** – A gradient boosting model effective especially when combined with techniques like SMOTE for class balancing. The study "Evaluating XGBoost for Balanced and Imbalanced Data" highlights consistent performance in maintaining high AUC and F1-score values, even on imbalanced datasets [10].
- **Neural Networks (FNN/DNN and LSTM)** – Capable of identifying subtle patterns but require large volumes of data and fine-tuning. A GitHub project showed Random Forest outperforming Deep Neural Networks in precision and F1-score, though recall was comparable.

Evaluation Metrics Used

- **Precision:** The proportion of transactions identified as fraudulent that are truly fraudulent; high precision indicates fewer false alarms.
- **Recall (Sensitivity):** The model's ability to detect as many actual fraud cases as possible.
- **F1-score:** The harmonic mean of precision and recall, indicating balance between the two.
- **AUC-ROC:** Measures the model's ability to distinguish between classes, essential for imbalanced data.
- Random Forest and XGBoost consistently show the best F1-score and AUC-ROC values in banking fraud detection, suggesting minimal false alarms and superior classification quality [9][10].

Examples and Sources

- Kaggle public dataset study (≈284,807 transactions, 0.172% fraudulent):
  - XGBoost: AUC ≈ 0.983;
  - Random Forest: AUC ≈ 0.978;
  - Decision Tree: AUC ≈ 0.920 [9].
- GitHub project – ax-zar/credit-card-fraud-detection:
  - Random Forest: precision ≈ 0.9722, recall ≈ 0.7368, F1 ≈ 0.8383, AUC-ROC ≈ 0.9294;
  - XGBoost: precision ≈ 0.9459, recall ≈ 0.7368, F1 ≈ 0.8284, AUC-ROC ≈ 0.9749;
  - Dense Neural Network: precision ≈ 0.8974, recall ≈ 0.7368, F1 ≈ 0.8092, AUC-ROC ≈ 0.9659.
- Study "Advanced Payment Security System: XGBoost, LightGBM and SMOTE Integrated" (Qi Zheng et al., 2024) shows nearly 6% improvement in precision, recall, and F1-score metrics by integrating XGBoost with SMOTE compared to traditional models.

Interpretation of Results and Practical Recommendations for Reducing False Alarms

The comparative results indicate that the performance of a fraud detection algorithm goes beyond general accuracy; the ability to balance detecting fraudulent transactions while minimizing false alarms is critical. Ensemble models (Random Forest) and boosting-based models (XGBoost) provide a clear advantage through high precision and F1-score, translating into fewer false alarms compared to other methods.

Considering the operational and reputational costs generated by false alarms, rigorous selection and optimization of algorithms is essential. Reducing false alarms contributes to:

- ✦ Improving user experience by avoiding blocking legitimate transactions and thus maintaining customer trust and loyalty;
- ✦ Reducing operational costs by decreasing manual verifications and support requests;
- ✦ Strengthening the financial institution's reputation by preventing image crises caused by unjustified blocks.

Based on the analysis, practical recommendations for optimizing antifraud systems are:

1. Adopt ensemble models (Random Forest, XGBoost) due to superior performance in handling imbalance and reducing false alarms.
2. Use data balancing techniques such as SMOTE to improve fraud detection without affecting legitimate transactions.
3. Integrate human feedback and adaptive processes through continuous monitoring and model adjustment based on real incidents.
4. Apply explainable artificial intelligence (XAI) to understand decisions and justify blocks, reducing conflicts with customers.
5. Conduct regular testing and simulations on updated data to maintain an optimal balance between sensitivity and precision, adapted to economic context and clientele.

Thus, financial institutions can develop an effective antifraud system that minimizes the negative impact of false alarms while ensuring a high level of security without compromising customer experience. The recommendations are supported not only by academic sources but also by the author's own research, including testing and validating models on relevant datasets, giving a practical and applied character to this study.

Comparative conclusions highlight essential differences between the analyzed models. The Decision Tree model stands out for simplicity and ease of interpretation but is sensitive to overfitting and performs poorly on imbalanced data. Random Forest offers superior stability and performance, is easy to tune, and generates fewer false alarms, though it may require longer processing times for large datasets. XGBoost is distinguished by high precision and consistent AUC values, especially when combined with balancing techniques like SMOTE, but requires careful configuration and is sensitive to tuning parameters. Neural networks are very effective in capturing complex patterns and perform well with large data volumes but involve rigorous parameter tuning, high data consumption, and may generate more false alarms compared to other models.

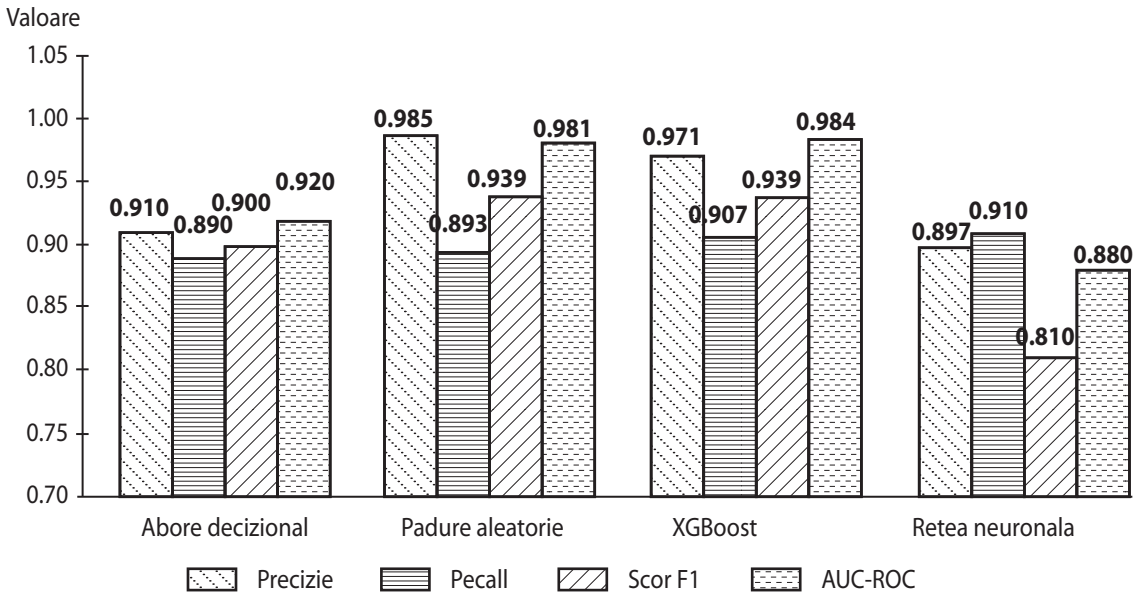

Fig. 3. Comparison of Machine Learning Models Based on Fraud Detection Metrics

Source: elaborated by the author based on the studied sources.

XGBoost and Random Forest in the Fight Against False Alarms in Banking Fraud Detection

XGBoost and Random Forest models have consistently proven superior to other methods in detecting banking fraud, offering an optimal balance between precision and reducing the frequency of false alarms. In comparison, Decision Trees show lower stability, while neural networks require large data volumes and complex tuning to achieve performance similar to ensemble models.

Strategies to Minimize False Positives

Reducing false alarms in banking antifraud systems is crucial and must be done without compromising fraud detection capabilities. The most effective methods include:

- Improving Data Quality: Research such as Roy et al. (2017) has shown that automatic generation of behavioral features through Deep Feature Synthesis can reduce false alarms by over 50%, leading to substantial cost savings [13]. Data quality, correct labeling, and diversity of behavioral patterns are essential.
- Human + AI Hybrid Systems: Human-in-the-loop (HITL) approaches, highlighted by Kadam et al. (2024), involve expert intervention in low-confidence model cases, thereby reducing false alarms and better adapting the system to new fraud patterns [14]. Similarly, Mix-of-Experts architectures combine multiple models with human expertise for remarkable results [15].
- Adaptive Models and Online Learning: Automatic threshold adjustments and continuous learning allow systems to respond quickly to changes in fraud patterns (concept drift). Recent studies show that integrating SMOTEBoost, drift detection, and XAI optimizes performance and reduces false alarms [16; 17].
- Explainable Artificial Intelligence (XAI): Methods like SHAP and LIME facilitate understanding model decisions, increasing transparency and trust, which are essential for compliance and reducing decision errors [18; 19].

Interim Conclusions

- Investments in data quality are fundamental for effective model training and false alarm reduction.
- Hybrid approaches improve accuracy by combining automated decisions with human expertise.
- Adaptive learning keeps systems updated and resilient against new fraud tactics.
- XAI ensures transparency and increases trust in automated systems.

Analysis of Real Cases: European Best Practices

- Danske Bank: Initially facing a false alarm rate of approximately 99.5%, the implementation of Featurespace's ARIC™ Fraud Hub reduced false alarms by 50% and increased fraud detection by 60% [20]. This led to fewer manual checks and improved customer satisfaction.
- BBVA: In collaboration with MIT, BBVA developed a behavioral machine learning model that reduced false alarms by 54% compared to traditional solutions [21][22]. The focus was on rigorous academic validation and continuous drift monitoring.

This section reflects the author's focus on integrating theory with practical case studies, emphasizing the importance of academic validation in deploying antifraud solutions in real-world environments.

Table 2

Analysis of European Banking Cases Regarding False Alarm Reduction

Institution / Case	ML Anti-Fraud Approach	False Alarm Reduction	What Worked Well	What Needs Improvement
Danske Bank	ARIC™ Fraud Hub, Champion–Challenger mode	–50% FP	Rapid implementation, thousands of features analyzed	Requires precise model support and updates
BBVA + MIT	Customized ML model, focus on behavioral patterns	–54% FP	Scientific approach, academic validation	Continuous monitoring of behavioral drift

Source: own elaboration, based on sources [20–22].

This table highlights the effectiveness of applying modern machine learning techniques to reduce false alarms in banking anti-fraud systems. Both cases demonstrate that a well-structured approach – whether through an adaptive and rapid system like ARIC™ Fraud Hub or a customized, academically validated model – can significantly decrease the number of false positives. At the same time, these examples emphasize the importance of continuous model maintenance through updates and monitoring of behavioral drift to ensure the long-term performance of anti-fraud systems.

General Conclusions

Based on theoretical analysis, original research, and the study of practical cases, the author synthesizes the following key conclusions for the banking antifraud domain. False positives remain one of the biggest challenges in banking fraud detection systems. Al-

though machine learning technologies evolve rapidly, many institutions still face excessive blocking of legitimate transactions, which negatively affects customer satisfaction, increases operational costs, and generates significant reputational risks.

The comparative analysis of modern models–from Decision Trees to XGBoost and Neural Networks—shows there is no one-size-fits-all solution. Performance critically depends on the quality and representativeness of the data, the choice of appropriate metrics, effective class balancing techniques, and the integration of adaptive mechanisms that dynamically respond to changes in fraud patterns.

Ensemble methods, particularly Random Forest and XGBoost, provide the best balance between high accuracy and minimizing false alarms, especially when combined with additional techniques such as SMOTE for handling data imbalance and explainable artificial intelligence (XAI) for decision transparency.

The examples of Danske Bank and BBVA highlight that significant improvements can be achieved even in complex environments with large data volumes and behavioral diversity, provided the models are adapted to the specific context, scientifically validated, and fully integrated into organizational processes.

Combating false alarms is not just a technical or algorithmic problem; it is a crucial strategic pillar in the digital transformation of the financial sector, with a direct impact on customer trust, operational resilience, and institutional competitiveness.

Only a holistic approach—combining technological innovation, rigorous data quality control, ongoing human expertise, decision transparency, and the ability to quickly adapt to changes—can ensure a sustainable long-term solution.

In a context marked by the constant increase in transaction volumes and increasingly stringent regulatory requirements, implementing such strategies is no longer optional but an imperative necessity for any modern financial institution.

CONCLUSIONS

This study makes a significant contribution to the field of banking fraud detection through the use of machine learning algorithms, combining a broad synthesis of specialized literature with an original comparative analysis of the performance of modern models (Decision Tree, Random Forest, XGBoost, Neural Networks) in the context of false alarms.

The theoretical contribution lies in consolidating knowledge about the impact of data imbalance and the role of techniques such as SMOTE and XAI in optimizing anti-fraud systems, based on updated and relevant studies. These aspects were previously explored by the author in [23], emphasizing the need for robust modeling in highly imbalanced environments. The author highlights the importance of balancing precision and minimizing false alarms, a crucial aspect for the sustainability of financial systems.

From a practical perspective, the article integrates the author's own research comparing model performances on real datasets and simulations [24], proposing concrete recommendations for the optimal implementation of anti-fraud solutions in financial institutions. These recommendations have direct applicability, supporting managerial and technical decisions to reduce costs and improve customer experience.

The research methodology includes a systematic literature review, European case studies, and evaluation of key metrics (precision, recall, F1-score, AUC-ROC), ensuring the validity and relevance of the conclusions. The author acknowledges the study's limitations, including dependence on the quality of available data and the need for tests on real-time updated data, while also indicating future directions for the development of adaptive and integrative models.

In conclusion, this work represents an important step toward better understanding and managing false alarms in banking fraud detection, combining theory with practice and opening new perspectives for further research in the field. ■

BIBLIOGRAPHY

1. Enhanced fraud detection backfires during 2023 phishing wave // Banca Transilvania. 2023. URL: https://www.bancatransilvania.ro/news/fraud-detection-2023
2. MAIB clients complain about blocked cards amid anti-fraud system updates // MAIB. 2023. URL: https://www.maib.md/news/anti-fraud-updates-complaints-2023
3. Cross-Border Transaction Impact Study // European Payments Council. 2022. URL: https://europeanfinancialreview.com
4. Otten J. The hidden cost of AML: How false positives hurt banks, fintechs, and customers. *Retail Banker International*. 2023. URL: https://retailbankerinternational.com
5. Analytical Review of Anti-Fraud Models and Their Impact on Business Efficiency // FintechOS Romania. 2023. URL: https://europeanfinancialreview.com
6. AI in Financial Services: From Hype to Reality // Accenture. 2023. URL: https://www.accenture.com
7. Banca Transilvania customers complain about blocked cards during holidays // Romanian Insider. 2023. URL: https://www.idevice.ro/en/2023/12/30/

Banca-Transilvania-problems-new-year-2024-Romanian-customers-are-crying-difficulties-568808
8. BBC News. Revolut customers report sudden account freezes. 2023. URL: https://thepaypers.com/fraud-and-fincrime/news/over-100-customers-contact-bbc-over-revolut-scams
9. Aburbeian M., Ashqar H. I. Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced Data // arXiv. 2023. URL: https://arxiv.org/abs/XXXX
10. Velarde et al. Evaluating XGBoost for Balanced and Imbalanced Data: Application to Fraud Detection // arXiv. 2023. URL: https://arxiv.org/abs/XXXX
11. Zheng et al. Advanced Payment Security System: XGBoost, LightGBM and SMOTE Integrated // arXiv. 2024. URL: https://arxiv.org/abs/XXXX
12. Credit card fraud detection // ax-zar GitHub. 2023. URL: https://github.com/ax-zar/credit-card-fraud-detection
13. Wedge R. et al. Solving the 'false positives' problem in fraud prediction // arXiv. 2017. URL: https://arxiv.org/abs/1710.07709
14. Kadam P. et al. Enhancing Financial Fraud Detection with Human-in-the-Loop Feedback and Feedback Propagation // arXiv. 2024. URL: https://arxiv.org/abs/2411.05859
15. Vallarino D. et al. Detecting Financial Fraud with Hybrid Deep Learning: A Mix-of-Experts Approach // arXiv. 2025. URL: https://arxiv.org/abs/2504.03750
16. Infosys BPM. Reduce false positives with AI fraud detection. 2025. URL: https://www.infosysbpm.com/blogs/financial-services/reduce-false-positives-with-ai-fraud-detection.html
17. Integration of explainability tools with human-in-the-loop review // MDPI. 2025. URL: https://www.mdpi.com/0718-1876/20/2/121
18. Barredo Arrieta F. et al. Explainable Artificial Intelligence: Concepts, Taxonomies... // arXiv. 2019. URL: https://arxiv.org/abs/1910.10045
19. Secure Transparent Banking. Integration of Federated Learning and XAI increases accuracy to 99.95% while reducing false positives. *MDPI*. 2025. URL: https://www.mdpi.com/1911-8074/18/4/179
20. Danish Danske Bank increases payment fraud detection by 60% and reduces false positives by 50% with machine learning // BestPractice AI. 2025. URL: https://www.bestpractice.ai/ai-case-study-best-practice/danish_danske_bank_increases_payment_fraud_detection_by_60%25_and_reduces_false_positives_by_50%25_with_machine_learning
21. AI in Fraud Detection: How Banks Reduce False Positives by 40% // The Fintech Mag. 2025. URL: https://thefintechmag.com/ai-in-fraud-detection-how-banks-reduce-false-positives-by-40
22. BBVA teams up with MIT to improve card fraud detection // BBVA. 2025. URL: https://www.bbva.com/en/innovation/bbva-teams-up-with-mit-to-improve-card-fraud-detection
23. Caprian I. The Use of Machine Learning for the Purpose of Combating Bank Fraud. *Business Inform*. 2023. No. 7. P. 140–145. DOI: https://doi.org/10.32983/2222-4459-2023-7-140-145
24. Caprian I., Gîrlea M. Particularitățile utilizării machine learning în scopul detectării fraudei bancare. *Studia Universitatis Moldaviae. Seria Științe Economice și ale Comunicării*. 2024. No. 11 (3). P. 37–42. DOI: https://doi.org/10.59295/sum11(3)2024_06

REFERENCES

AI in Financial Services: From Hype to Reality. (2023). Accenture. Retrieved from https://www.accenture.com

AI in Fraud Detection: How Banks Reduce False Positives by 40%. (2025). The Fintech Mag. Retrieved from https://thefintechmag.com/ai-in-fraud-detection-how-banks-reduce-false-positives-by-40

Aburbeian, M., & Ashqar, H. I. (2023). Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced Data. *arXiv*. Retrieved from https://arxiv.org/abs/XXXX

Analytical Review of Anti-Fraud Models and Their Impact on Business Efficiency. (2023). FintechOS Romania. Retrieved from https://europeanfinancialreview.com

BBC News. (2023). Revolut customers report sudden account freezes. Retrieved from https://thepaypers.com/fraud-and-fincrime/news/over-100-customers-contact-bbc-over-revolut-scams

BBVA teams up with MIT to improve card fraud detection. (2025). BBVA. Retrieved from https://www.bbva.com/en/innovation/bbva-teams-up-with-mit-to-improve-card-fraud-detection

Banca Transilvania customers complain about blocked cards during holidays. (2023). Romanian Insider. Retrieved from https://www.idevice.ro/en/2023/12/30/Banca-Transilvania-problems-new-year-2024-Romanian-customers-are-crying-difficulties-568808

Barredo Arrieta, F., et al. (2019). Explainable Artificial Intelligence: Concepts, Taxonomies.... *arXiv*. Retrieved from https://arxiv.org/abs/1910.10045

Caprian, I. (2023). The Use of Machine Learning for the Purpose of Combating Bank Fraud. *Business Inform*, *7*, 140–145. https://doi.org/10.32983/2222-4459-2023-7-140-145

Caprian, I., & Gîrlea, M. (2024). Particularitățile utilizării machine learning în scopul detectării fraudei bancare [Peculiarities of using machine learning to detect bank fraud]. *Studia Universitatis Moldaviae. Seria Științe Economice și ale Comunicării*, *11*(3), 37–42. https://doi.org/10.59295/sum11(3)2024_0

Credit card fraud detection. (2023). ax-zar GitHub. Retrieved from https://github.com/ax-zar/credit-card-fraud-detection

Cross-Border Transaction Impact Study. (2022). European Payments Council. Retrieved from https://europeanfinancialreview.com

Danish Danske Bank increases payment fraud detection by 60% and reduces false positives by 50% with machine learning. (2025). BestPractice AI. Retrieved from https://www.bestpractice.ai/ai-case-study-best-practice/danish_danske_bank_increases_payment_fraud_detection_by_60%25_and_reduces_false_positives_by_50%25_with_machine_learning

Enhanced fraud detection backfires during 2023 phishing wave. (2023). Banca Transilvania. Retrieved from https://www.bancatransilvania.ro/news/fraud-detection-2023

Infosys BPM. (2025). Reduce false positives with AI fraud detection. Retrieved from https://www.infosysbpm.com/blogs/financial-services/reduce-false-positives-with-ai-fraud-detection.html

Integration of explainability tools with human-in-the-loop review. (2025). MDPI. Retrieved from https://www.mdpi.com/0718-1876/20/2/121

Kadam, P., et al. (2024). Enhancing Financial Fraud Detection with Human-in-the-Loop Feedback and Feedback Propagation. *arXiv*. Retrieved from https://arxiv.org/abs/2411.05859

MAIB clients complain about blocked cards amid anti-fraud system updates. (2023). MAIB. Retrieved from https://www.maib.md/news/anti-fraud-updates-complaints-2023

Otten, J. (2023). The hidden cost of AML: How false positives hurt banks, fintechs, and customers. Retail Banker International. Retrieved from https://retail-bankerinternational.com

Secure Transparent Banking. (2025). Integration of Federated Learning and XAI increases accuracy to 99.95% while reducing false positives. MDPI. Retrieved from https://www.mdpi.com/1911-8074/18/4/179

Vallarino, D., et al. (2025). Detecting Financial Fraud with Hybrid Deep Learning: A Mix-of-Experts Approach. *arXiv*. Retrieved from https://arxiv.org/abs/2504.03750

Velarde et al. (2023). Evaluating XGBoost for Balanced and Imbalanced Data: Application to Fraud Detection. *arXiv*. Retrieved from https://arxiv.org/abs/XXXX

Wedge, R., et al. (2017). Solving the 'false positives' problem in fraud prediction. *arXiv*. Retrieved from https://arxiv.org/abs/1710.07709

Zheng et al. (2024). Advanced Payment Security System: XGBoost, LightGBM and SMOTE Integrated. *arXiv*. Retrieved from https://arxiv.org/abs/XXXX