BMBAD BJESR Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches

Benchmarking Machine Learning Models for Transaction-Level Anti-Money Laundering under Extreme Class Imbalance

Aşırı Sınıf Dengesizliği Altında İşlem Düzeyinde Kara Para Aklama Tespiti için Makine Öğrenmesi Modellerinin Kıyaslanması

[1]Ümit YILMAZ

[1]Balıkesir Üniversitesi, Bigadiç Meslek Yüksekokulu, Bigadiç/Balıkesir, Türkiye

[1]umityilmaz@balikesir.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO

Article history

Received : 24 December 2025
Accepted : 5 March 2026

Keywords:

Anti-Money Laundering, Financial Fraud Detection, Machine Learning, Class Imbalance, Transaction Classification

ABSTRACT

This study presents a comparative evaluation of supervised machine learning methods for transaction-level anti-money laundering detection under extreme class imbalance. Experiments are conducted on the Lower Illicit–Small subset of the IBM Transactions for anti-money laundering dataset, which includes nearly seven million transactions with a laundering prevalence of approximately 0.05%. Logistic Regression, Random Forest, XGBoost, and CatBoost models are implemented using model-specific preprocessing and hyperparameter tuning strategies. Model performance is assessed on an independent test set using accuracy, precision, recall, F1-score, ROC-AUC, and balanced accuracy metrics. The results reveal clear precision–recall trade-offs among the models. While Logistic Regression reflects an inclusive detection approach with the highest recall and balanced accuracy, CatBoost supports more selective alarm strategies aimed at minimizing false positives, driven by its higher precision and ROC–AUC performance. Overall, the findings highlight the importance of metric-driven model selection and careful operating-point design in practical anti-money laundering screening systems.

© 2026 Bandirma Onyedi Eylul University, Faculty of Engineering and Natural Science. Published by Dergi Park. All rights reserved.

MAKALE BİLGİSİ

Makale Tarihleri

Gönderim : 24 Aralık 2025
Kabul : 5 Mart 2026

Anahtar Kelimeler:

Kara Para Aklama, Finansal Dolandırıcılık Tespiti, Makine Öğrenmesi, Sınıf Dengesizliği, İşlem Sınıflandırması

ÖZET

Bu çalışma, aşırı sınıf dengesizliği koşulları altında işlem düzeyinde kara para aklama tespitine yönelik denetimli makine öğrenmesi yöntemlerinin karşılaştırmalı bir değerlendirmesini sunmaktadır. Deneyler, yaklaşık yedi milyon işlemden oluşan ve kara para aklama oranı yaklaşık %0,05 olan IBM Transactions for Anti-Money Laundering veri setinin Lower Illicit–Small alt kümesi üzerinde gerçekleştirilmiştir. Lojistik Regresyon, Rastgele Orman, XGBoost ve CatBoost modelleri, modele özgü ön işleme adımları ve hiperparametre ayarlamaları kullanılarak uygulanmıştır. Model performansı, bağımsız bir test kümesi üzerinde doğruluk, kesinlik, duyarlılık, F1-skoru, ROC-AUC ve dengeli doğruluk ölçütleri kullanılarak değerlendirilmiştir. Bulgular, modeller arasında belirgin kesinlik–duyarlılık dengeleri bulunduğunu göstermektedir. Lojistik Regresyon, en yüksek duyarlılık ve dengeli doğruluk değerleriyle kapsamlı bir tespit yaklaşımını yansıtırken, CatBoost daha yüksek kesinlik ve ROC-AUC alan performansı ile yanlış pozitifleri minimize etmeye yönelik, daha seçici alarm stratejilerini desteklemektedir. Sonuçlar, uygulamalı kara para aklama tarama sistemlerinde ölçüt temelli model seçiminin ve uygun karar eşiklerinin belirlenmesinin önemini ortaya koymaktadır.

© 2026 Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi. Dergi Park tarafından yayımlanmaktadır. Tüm Hakları Saklıdır.

ORCID ID: 0000-0003-4268-8598

1. INTRODUCTION

Money laundering, the practice of moving illicit funds through financial systems to disguise their origin, remains a pervasive global concern. According to the United Nations Office on Drugs and Crime (UNODC), it is estimated that roughly 2–5% of global GDP is laundered annually, corresponding to an estimated USD 800 billion to USD 2 trillion in illicit funds each year [1]. Beyond the direct financial losses, money laundering facilitates a wide range of criminal activities, including cyber-fraud and human trafficking, by enabling perpetrators to enjoy illicit profits while evading detection. Financial institutions are at the forefront of anti-money laundering (AML) efforts, yet traditional rule-based and manual monitoring approaches struggle to address increasingly sophisticated and large-scale laundering schemes [2, 3]. These conventional systems not only generate an overwhelming number of false positives, which flag benign transactions as suspicious, but also fail to detect subtle laundering patterns, thereby undermining enforcement effectiveness [4, 5].

Machine learning (ML) presents a promising avenue for enhancing AML systems by automatically identifying suspicious behavioral patterns within large-scale financial data [6-8]. However, a significant challenge in building ML models for money laundering detection lies in the scarcity of labeled datasets. Real financial transaction records are proprietary and constrained by strict privacy regulations, while confirmed laundering cases are rare and frequently underreported due to undetected activities. To address this limitation, researchers have increasingly turned to synthetic datasets that emulate real-world financial systems [9]. In this study, the IBM Transactions for AML dataset is employed, a publicly available synthetic dataset generated through multi-agent simulation of financial transactions. This dataset provides a realistic testbed for AML model evaluation, encompassing millions of records of banking transactions (such as transfers and payments), with only a minute fraction labeled as illicit. Specifically, the “Lower Illicit – Small” subset is utilized, which reflects real-world AML conditions through millions of transactions and an extreme class imbalance, where illicit activities constitute only a minute fraction of the data.

The objective of this study is to evaluate and compare the performance of supervised ML classification algorithms in detecting money laundering transactions within the IBM dataset. The central research question addressed is: which classification methods and preprocessing strategies are most effective at identifying rare illicit transactions while minimizing false positives? The work specifically tackles key challenges such as handling extreme class imbalance, processing heterogeneous features (including transaction amounts, currencies, account identifiers, and payment types), and selecting appropriate evaluation metrics tailored to imbalanced classification. While prior studies have applied ML to AML detection across various data modalities; including cryptocurrency traces [2, 10], e-wallet streams [4], and institutional transaction records [3]; a systematic comparison of classical and gradient-boosted classifiers on a large-scale, publicly available synthetic banking dataset under extreme class imbalance, with explicit analysis of precision–recall trade-offs and their operational implications, remains underexplored. This study addresses that gap by providing a reproducible benchmark that connects modeling choices to deployment considerations in practical AML screening systems.

To this end, a systematic analysis is conducted using four machine learning models: (1) Random Forest (RF), (2) Logistic Regression (LR), (3) XGBoost, and (4) CatBoost, a gradient boosting method optimized for categorical features. Data preprocessing pipelines were designed for each model, including appropriate scaling and encoding procedures, and hyperparameters were optimized through cross-validation. Model performance was assessed on an independent test set using a comprehensive set of metrics, namely accuracy, precision, recall, F1-score, Receiver Operating Characteristic-Area Under the Curve (ROC-AUC), and balanced accuracy. Against this backdrop, prior studies have explored a wide range of data sources and modeling strategies for AML detection.

2. RELATED WORK

In parallel with these methodological choices, recent AML research has diversified across data modalities, including bank transactions, on-chain traces, and suspicious activity report (SAR) filings, as well as across learning paradigms such as supervised, unsupervised, graph-based, and hybrid approaches, with increasing emphasis on class imbalance, threshold calibration, and explainability. On cryptocurrency data, Venčkauskas, et al. [2] introduced a value-driven transactional tracking approach that combines de-anonymized on-chain data with structured training pipelines and compliance-oriented advisory components, demonstrating the effectiveness of tree-based ensemble models for forensic analysis. Similarly, Salahi, et al. [10] reported that ensemble-based classifiers are well suited for distinguishing licit and illicit blockchain transactions. Extending beyond static representations, Wan and Li [11] proposed a spatiotemporal framework integrating dynamic graph convolution with recurrent modeling to better capture evolving transaction patterns, while E and S. S [12] formulated blockchain tumbler detection as a multi-class problem, further reinforcing the robustness of tree-based ensembles on feature-rich on-chain data.

On bank and fintech transaction streams, a prominent line of work addresses extreme class imbalance and operating-point selection. Using e-wallet transaction data, Doddamani, et al. [4] investigated resampling strategies and threshold calibration to improve minority-class detection. Deep learning models have also been explored: Konstantinidis and Gegov [6] employed neural architectures combined with explainability techniques to enhance transparency, whereas Kute, et al. [7] proposed an explainable convolutional framework emphasizing recall-

oriented alerting for compliance settings. Complementing supervised classifiers with anomaly detection, Sharma, et al. [5] introduced a hybrid pipeline that integrates weak supervision, resampling, classical learners, neural networks, and isolation-based methods to support workload-aware alert triage.

Graph-centric and unsupervised approaches have been deployed to surface atypical accounts and transaction structures that may evade per-transaction models. Irshad, et al. [13] combined graph centrality measures with transactional attributes to identify suspicious patterns, while typology-aware features were shown to enhance classical models even without fully fledged graph neural networks [14]. Beyond transaction-level detection, broader AML intelligence tasks have also been explored. Gandhi, et al. [15] analyzed FinCEN SAR data to capture spatial–temporal patterns in regulatory reporting, whereas Lokanan [16] examined macro-level AML risk using institutional and governance indicators. From an operational perspective, Alexandre and Balsa [3] integrated client profiling and rule-based reasoning to prioritize high-risk entities at scale.

A parallel strand emphasizes criminology-grounded features, explainable artificial intelligence, and fairness considerations. Ramadhan [8] demonstrated that integrating criminological theory with explainability and fairness auditing can improve both model transparency and decision equity. Collectively, these studies indicate that effective AML detection requires careful alignment between modeling strategies, evaluation priorities, and operational constraints, particularly under conditions of extreme class imbalance.

Within this landscape, the present study benchmarks RF, LR, XGBoost, and CatBoost on the IBM AML dataset and documents the precision–recall trade-offs that arise under extreme imbalance, where LR favoring coverage (high recall) and CatBoost favors precision, thereby motivating tiered or rank-based workflows and explicit operating-point selection. By doing so, the study provides empirical guidance for model selection and metric prioritization in transaction-level AML screening. Unlike prior work that often focuses on a single model family or evaluates on proprietary or small-scale datasets, the present study contributes a multi-model comparison on a large, publicly accessible benchmark under realistic prevalence conditions, with explicit attention to operational implications. The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the dataset, modeling approach, and evaluation metrics. Section 4 presents and discusses the experimental results. Section 5 concludes the paper and outlines directions for future research.

The present study contributes to the literature in three ways. First, it provides a rigorous multi-model benchmark comparing Logistic Regression, Random Forest, XGBoost, and CatBoost on a large-scale (approx. 7 million transactions), publicly available synthetic banking dataset (IBM AML) under extreme class imbalance (0.05% prevalence), yielding reproducible and directly comparable results. Second, it offers an empirical analysis of precision–recall trade-offs and their operational implications, demonstrating that no single model dominates across all metrics and that model selection must be driven by institutional cost priorities (coverage-first vs. precision-first workflows). Third, it presents evidence that CatBoost's native handling of high-cardinality categorical identifiers yields the strongest ROC-AUC (0.9648) and highest precision, while Logistic Regression achieves the highest balanced accuracy (0.8787), establishing clear, metric-driven baselines for practical AML screening system design.

3. METHODOLOGY

3.1. Data Description

In this study, the "Lower Illicit, Small" subset of the IBM Transactions for AML dataset was employed [17]. The corpus was generated synthetically within a virtual financial ecosystem populated by individuals, companies, and banks. Agents conduct a wide range of financial interactions through bank accounts, while a small subset engages in criminal activity. The data generator traces illicit funds across multiple transactions and assigns ground-truth labels to individual records as laundering or legitimate [9].

The laundering process was modelled end-to-end through placement, layering, and integration stages. Unlike real institutional logs that provide only partial visibility, the synthetic setting represents a system-wide view of transactions across institutions, thereby enabling the development and assessment of detection models under broad network coverage.

Each record contains eleven attributes. These include a transaction timestamp; sender and receiver bank codes identified as From Bank and To Bank; sender and receiver account identifiers; the transacted amount and its currency; a categorical transaction (cheque, wire transfer, or reinvestment); and the binary label Is Laundering, where the value one denotes laundering and the value zero denotes a legitimate transaction. In this subset, the sent and received amounts and currencies are identical, which effectively yields a single numerical amount feature. No additional temporal features were derived from the timestamp, and the engineering of such features is left as future work.

A stratified split of 80/20 was applied to the LI-Small subset, which contains approximately 7 million transactions, in order to create training and test sets while preserving the extreme rarity of positive cases in both partitions. Approximately 5.54 million transactions were assigned to the training set and 1.38 million to the test set. The overall prevalence of laundering in the LI-Small configuration is low, which corresponds approximately to one laundering transaction for every two thousand legitimate transactions. Measures were taken to avoid label leakage and to limit overfitting to the scarce positive instances [17].

3.2. Modeling and Hyperparameter Tuning

The four classifiers considered in this study instantiate complementary learning biases that are well-suited to tabular AML data under extreme class imbalance. These models were selected for three principal reasons. First, they span a broad spectrum of inductive biases, from the convex, probabilistic formulation of LR to the nonparametric ensemble averaging of RF and the sequential, boosting-based optimization of XGBoost and CatBoost, enabling a rigorous comparison of fundamentally different learning strategies on the same benchmark. Second, all four methods natively support class-weighted training or cost-sensitive objectives, which is a critical requirement when positive-class prevalence is approximately 0.05%. Third, they represent the most widely adopted baselines in tabular fraud and AML detection research [2-5, 7, 10], making the results directly comparable to prior work and practically relevant to financial institutions evaluating off-the-shelf solutions. The modeling objective for all methods may be cast as empirical risk minimization on a surrogate classification loss with class-weighted contributions to reflect the disproportionate cost of missed laundering relative to false alarms. Threshold-dependent metrics were optimized indirectly by searching over model hyperparameters while monitoring validation F1.

- **Logistic Regression (LR):** LR minimizes a convex, regularized logistic loss, yielding a unique global optimum and stable convergence due to the objective's convexity. In practice, it serves as a strong baseline classifier with well-calibrated probabilistic outputs, making it suitable for imbalanced settings where calibrated scores enable post-training threshold adjustment [18]. Regularization is used to control model complexity: an L2 (ridge) penalty shrinks coefficients to reduce variance and stabilize the model, whereas an L1 (lasso) penalty induces sparsity by driving some coefficients exactly to zero. The latter effectively performs feature selection in high-dimensional one-hot encoded feature spaces [19]. Incorporating class weights in LR's loss function is equivalent to cost-sensitive learning – it penalizes misclassification of minority-class examples more heavily, thereby shifting the decision boundary to favor recall of the positive class [20]. Under mild model assumptions, LR's output probabilities tend to be well-calibrated; thus one can adjust the classification threshold to meet operational precision/recall targets without retraining the model [18].
- **Random Forest (RF):** RF is a nonparametric ensemble method that builds multiple decision trees on bootstrapped samples and aggregates their predictions to reduce variance [21]. This variance-reducing strategy yields a robust model with improved accuracy, often outperforming single decision-tree classifiers on complex tasks [22]. The nonparametric nature of RF means it makes no distribution assumptions and can learn highly nonlinear decision boundaries from data [23]. Its model complexity and the bias–variance trade-off are governed by hyperparameters such as the number of trees, tree depth, and minimum samples for splits/leaves. Increasing the number or depth of trees generally lowers bias but can raise variance, whereas restricting tree depth or requiring larger leaf sizes helps curb overfitting [21]. RF can also handle class imbalance by adjusting class weights in the splitting criterion, which increases sensitivity to minority classes [24]. Moreover, tree-based models are invariant to monotonic transformations of numeric features – scaling or log-transforming an input does not change an RF's splits or performance [25]. However, RF requires categorical variables to be encoded as numerical values (e.g., one-hot encoding) before training [26]. For very high-cardinality categories, specialized encoding techniques or algorithms that natively handle categoricals are often more effective [27].
- **XGBoost:** XGBoost is a gradient-boosting algorithm that builds an ensemble of additive decision trees to optimize a regularized logistic loss [28]. It includes several mechanisms to prevent overfitting: a small learning rate slows down each boosting step, stabilizing the stage-wise optimization and improving generalization; a limit on tree depth (max_depth) controls model complexity; and stochastic subsampling of rows and columns (subsample and colsample_bytree) injects randomness at each iteration to reduce variance [29-31]. For imbalanced datasets, the scale_pos_weight hyperparameter applies a cost-sensitive weight to the minority class, increasing the penalty for misclassifying minority instances [31, 32]. Additionally, XGBoost supports validation-based early stopping, which halts training when additional trees no longer improve validation performance, a strategy that helps prevent overfitting and aids in selecting the optimal number of boosting rounds [33].
- **CatBoost:** CatBoost employs ordered target statistics and oblivious trees as its base learners, which together help mitigate target leakage in categorical feature processing and ensure more stable, balanced splits [34]. This categorical handling effectively performs a type of regularized target encoding via random permutations, allowing CatBoost to leverage high-cardinality categorical identifiers without incurring the dimensionality explosion of one-hot encoding [35]. Model complexity is controlled through standard hyperparameters – such as tree depth, number of boosting iterations, and learning rate – much like other Gradient Boosting Decision Trees frameworks [36]. An L2 regularization penalty on leaf values ("l2_leaf_reg") is applied to discourage extreme leaf weight values, thereby reducing overfitting [37]. Furthermore, similar to XGBoost, CatBoost supports an early stopping mechanism as a form of structural risk control to halt training when validation performance stops improving [38]. In class-imbalanced scenarios where recall of minority classes is critical, one can also utilize class weighting to bias the model toward those classes and improve sensitivity [39].

Hyperparameters for all four models were tuned via RandomizedSearchCV. To manage computational overhead, the search was run on a stratified 20% subsample of the training data with two-fold stratified cross-validation, optimizing the F1-score. For CatBoost, each trial additionally used an internal 80/20 training–validation split as an evaluation set, with early stopping applied after 50 consecutive iterations without improvement. After selection, the best CatBoost configuration was refit on the full training data using a 15% validation split and the same early-stopping criterion; XGBoost followed the same refit protocol with early stopping, whereas RF and LR were refit on the full training data without early stopping.
The hyperparameter ranges were selected based on literature and practical considerations and shown in Table 1. After the random search, the optimal hyperparameter configurations were identified by mean F1-score across folds. Final models were trained on the full training dataset using these settings. For CatBoost and XGBoost, early stopping was again used, reserving 15 % of the training set for validation. RF and LR were trained on the entire training data without early stopping.

Tablo 1. Hyperparameter ranges for each model.

Model	Hyperparameter	Range
RF	Number of Trees	100, 200, 300, 500
	Maximum Depth	None, 10, 20, 30
	Minimum Samples per Split	2, 5, 10
	Minimum Samples per Leaf	1, 2, 4
LR	Regularization Strength (C)	0.01, 0.1, 1, 10
	Penalty	L1, L2
XGBoost	Number of Estimators	100, 200, 300
	Maximum Depth	3, 5, 7, 9
	Learning Rate	0.01, 0.05, 0.1, 0.2
	Subsample	0.6, 0.8, 1.0
	Column subsample per tree	0.6, 0.8, 1.0
CatBoost	Number of Boosting Iterations	200, 500, 800
	Tree Depth	4, 6, 8, 10
	Learning Rate	0.01, 0.05, 0.1
	L2 Leaf Regularization	1, 3, 5, 7

3.3. Evaluation Metrics

Model performance was quantified using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Balanced Accuracy in this order. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy measures the proportion of correct predictions over all instances, as given in Equation 1 [40].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision (positive predictive value) captures the correctness of positive predictions, as defined in Equation 2, whereas Recall (true positive rate, TPR) captures the completeness of positive predictions, as defined in Equation 3 [41].

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-score represents the harmonic mean of precision and recall, offering a balanced evaluation metric in situations where false positives and false negatives must be jointly considered, as defined in Equation 4. This measure is especially suitable for imbalanced classification problems [42].

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

The ROC–AUC score is derived from the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across different threshold settings, where the ROC is formally defined by TPR and FPR as given in Equations 5–6 [43]. Higher AUC values signify better discrimination between classes [44]. The AUC is the integral of TPR with respect to FPR over $[0, 1]$, given in Equation 7 [45, 46].

$$TPR = Recall = Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

$$ROC - AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{7}$$

To address class imbalance explicitly, Balanced Accuracy averages sensitivity and specificity. Specificity, or the true negative rate (TNR), is defined in Equation 8, and Balanced Accuracy is given in Equation 9 [47].

$$TNR = \frac{TN}{TN + FP} \tag{8}$$

$$Balanced\ Accuracy = \frac{TPR + TNR}{2} \tag{9}$$

4. RESULTS AND DISCUSSION

The best hyperparameter configurations identified during the search are reported in Table 2. The hyperparameter configurations in Table 2 indicate that the models adapted to the data representation in distinct ways. For RF, the selected number of trees, depth, and leaf criteria reflected a high-capacity ensemble designed to balance complexity and robustness. LR was optimized with strong regularization and an L2 penalty, suggesting that generalization was prioritized in the high-dimensional, sparse feature space. The configuration of XGBoost, which combined a moderate learning rate with a relatively deep structure and subsampling, was indicative of a staged learning process aimed at controlling overfitting. CatBoost was tuned with moderate iterations, shallow trees, and additional regularization, aligning with its strength in handling categorical features through ordered statistics and symmetric splits. All hyperparameters were selected based on mean F1-score during cross-validation, ensuring sensitivity to class imbalance.

Tablo 2. Best hyperparameters selected for each model.

Model	Hyperparameter	Best
RF	Number of Trees	200
	Maximum Depth	30
	Minimum Samples per Split	10
	Minimum Samples per Leaf	1
LR	Regularization Strength (C)	0.01
	Penalty	L2
XGBoost	Number of Estimators	300
	Maximum Depth	9
	Learning Rate	0.1
	Subsample	0.8
	Column subsample per tree	0.6
CatBoost	Number of Boosting Iterations	500
	Tree Depth	4
	Learning Rate	0.05
	L2 Leaf Regularization	5

The corresponding test-set performance across Accuracy, Precision, Recall, F1, ROC AUC, and Balanced Accuracy is summarized in Table 3. After training and tuning, the models were evaluated on the held-out test set comprising approximately 1.38 million transactions, where the true prevalence was about 0.05 %. Accuracy values remained very high, yet they conveyed limited information. In highly imbalanced settings, such values primarily reflected the dominance of the majority class and did not provide insight into the detection of rare laundering events. Balanced Accuracy is therefore emphasized alongside ROC-AUC and F1, as it averages sensitivity and specificity and is less sensitive to prevalence. RF achieved high Accuracy but exhibited negligible Recall and F1, showing that class weighting alone was insufficient when categorical signals were excluded. This behavior is consistent with the presence of high-cardinality categorical identifiers, which limit the effectiveness of conventional tree ensembles when no dedicated categorical encoding or target statistics are employed. LR produced a high ROC-AUC and very strong Recall, though Precision dropped considerably. This indicated that the operating point favored aggressive identification of positives, resulting in a large volume of false alarms. XGBoost yielded intermediate performance across metrics, with Recall and ROC-AUC higher than RF but lower than LR. CatBoost produced the strongest ROC-AUC and high Precision, but Recall remained limited at the default threshold, suggesting a conservative approach to labeling positives. Balanced Accuracy provided a prevalence-agnostic perspective on sensitivity and specificity, with values of 0.8787 for LR, 0.6398 for XGBoost, 0.5870 for CatBoost, and 0.5025 for RF, confirming substantial balance only for LR, moderate balance for XGBoost, modest balance for CatBoost, and chance-level performance for RF. While Balanced Accuracy provides a prevalence-agnostic summary of sensitivity and specificity, it may favor recall-heavy models under extreme imbalance. Accordingly, it is interpreted here as a complementary diagnostic metric rather than a standalone indicator of deployment readiness.

Figure 1 provides a consolidated view of the models' relative strengths. RF's near-perfect Accuracy coincided with very low Recall and F1, emphasizing its inability to capture laundering cases, and its Balanced Accuracy was close to chance level. LR demonstrated superior coverage of positives but at the cost of excessive false alarms, while attaining the highest Balanced Accuracy by combining strong sensitivity with reasonable specificity.

Tablo 3. Test set performance by model.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Balanced Accuracy
RF	0.9988	0.0042	0.0056	0.0048	0.7373	0.5025
LR	0.8009	0.0025	0.9565	0.0049	0.9164	0.8787
XGBoost	0.8993	0.0019	0.3801	0.0039	0.7844	0.6398
CatBoost	0.9996	0.8921	0.1739	0.2911	0.9648	0.5870

CatBoost showed the opposite tendency, strongly suppressing false positives while missing many laundering events, which was reflected in a modest Balanced Accuracy. XGBoost displayed a profile between these extremes, achieving intermediate Precision and Recall together with moderate Balanced Accuracy. These differences indicate that no single model can be considered dominant across all metrics and that threshold selection plays a critical role in aligning model behavior with operational objectives.

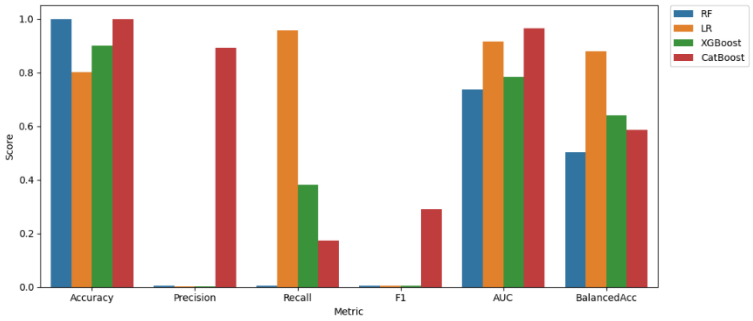


Figure 1. Test-set performance comparison across models.

The confusion matrices presented in Figure 2 clarifies these trade-offs. RF predictions were overwhelmingly concentrated in the majority class, producing almost no true positives, and Balanced Accuracy remained near chance as a result. LR yielded a marked rise in true positives together with a large block of false positives, yet it attained the highest Balanced Accuracy by pairing strong sensitivity with acceptable specificity. CatBoost demonstrated the opposite tendency, strongly suppressing false positives while allowing more false negatives, which was reflected in a modest Balanced Accuracy. XGBoost again provided an intermediate pattern, offering partial coverage of positives with limited precision and a correspondingly moderate Balanced Accuracy. These distributions reveal distinct cost profiles for each model and underscore the need to define organizational tolerance for missed cases versus false alerts.

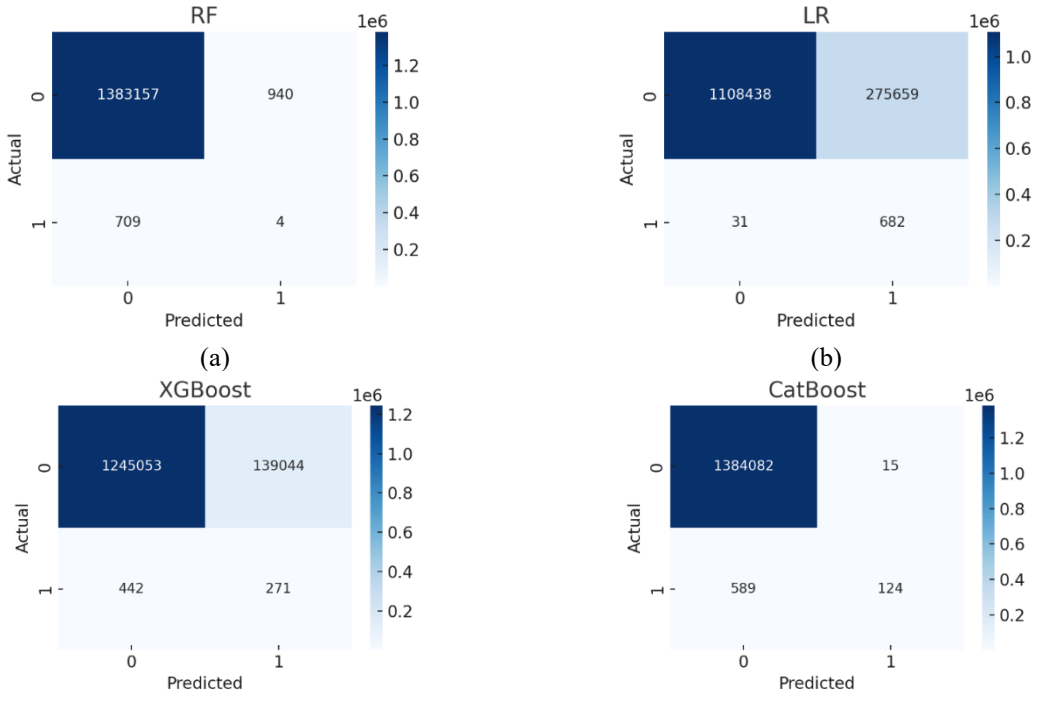


Figure 2. (a) RF, (b) LR, XGBoost, (d) CatBoost confusion matrices on the held-out test set.

The ROC curves in Figure 3 further illustrates ranking quality. The curves for CatBoost and LR extended closer to the upper left corner, yielding ROC-AUC values of approximately 0.9648 and 0.9164 respectively, confirming strong discriminative ability across thresholds. By contrast, RF and XGBoost occupied less favorable positions, reflecting weaker ranking capacity. It was observed, however, that a high ROC-AUC does not automatically translate into acceptable Precision at the default decision threshold, particularly under extreme imbalance. This finding suggests that precision–recall analysis and threshold sensitivity studies are indispensable for operational deployment. It should be emphasized that a high ROC-AUC reflects superior ranking capability across thresholds, rather than direct operational efficiency at any specific decision point. In highly imbalanced AML settings, strong ranking performance must therefore be interpreted in conjunction with precision–recall behavior and alert volume considerations.

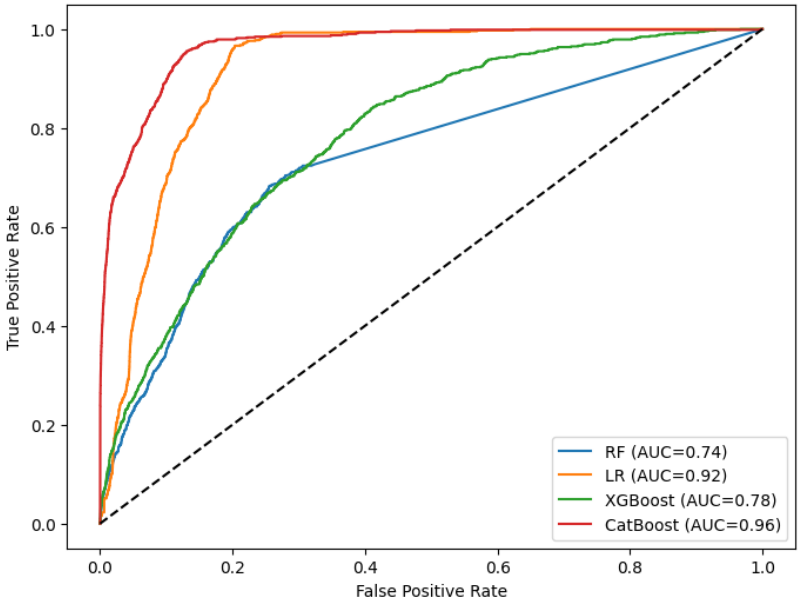


Figure 3. ROC curves on the held-out test set.

In summary, the comparative evaluation demonstrated that each model exhibited distinct advantages and limitations when applied to the detection of laundering transactions under extreme class imbalance. RF failed to raise sensitivity beyond chance levels despite near-perfect overall accuracy, while XGBoost provided only moderate improvement in recall and specificity. LR achieved the highest Balanced Accuracy by combining strong recall with acceptable specificity, though this came at the cost of excessive false alarms. CatBoost, in contrast, delivered the strongest ROC-AUC and the highest precision, indicating superior ranking capacity and conservative positive predictions that minimized false alarms. Taken together, these results suggest that CatBoost offered the most robust overall performance when the emphasis is placed on ranking quality and precision, while LR proved most effective in maximizing recall. The choice between these approaches therefore depends on organizational priorities, yet CatBoost emerges as a strong precision-oriented scorer with superior ranking capability, making it particularly suitable for conservative, precision-first AML workflows where false-positive control is critical. These findings are broadly consistent with, yet also meaningfully extend, prior work in the AML detection literature. The strong performance of tree-based ensemble and gradient-boosted models observed here aligns with results reported by Venčkauskas, et al. [2] and Salahi, et al. [10], who found that ensemble methods excel at separating illicit from licit transactions on cryptocurrency data. However, the present study reveals that on tabular banking data with categorical identifiers, vanilla RF underperforms severely at the default operating point due to its inability to leverage high-cardinality categoricals without dedicated encoding—a limitation not prominently discussed in those prior works. The superiority of CatBoost in precision and ROC-AUC is consistent with findings by Kundu, et al. [27] and König, et al. [35], who attribute CatBoost's advantage in high-cardinality categorical settings to its ordered target statistics. The strong recall of LR under class weighting corroborates observations by Doddamani, et al. [4] and Sun, et al. [20], who similarly found that cost-sensitive linear models with calibrated thresholds can achieve broad positive coverage at the cost of alert volume. The trade-off between precision and recall observed across all four models reinforces the finding of Sharma, et al. [5] and Konstantinidis and Gegov [6] that no single classifier dominates across all operational metrics in real-world AML settings, underscoring the importance of metric-driven selection aligned with institutional priorities. Unlike studies that employ deep learning [6, 7] or graph-based architectures [11, 13] for AML detection, which typically require more complex infrastructure and larger, richer datasets, the present study demonstrates that carefully tuned classical and gradient-boosted classifiers can achieve competitive discrimination (ROC-AUC up to 0.9648) on a large-scale publicly available benchmark, offering a practical and interpretable baseline for institutions with constrained computational resources.

5. CONCLUSION

An evaluation of four supervised learning methods—RF, LR, XGBoost, and CatBoost—was conducted on the IBM AML dataset under extreme class imbalance. While prior literature has examined individual model families or evaluated AML classifiers on proprietary and often small-scale datasets [2, 4, 6, 7, 10], a systematic multi-model benchmark on a large, publicly available banking transaction dataset with explicit analysis of precision–recall trade-offs and their operational implications remained underexplored. The present study addresses this gap by delivering a reproducible comparison that connects modeling choices directly to deployment considerations. The analysis underscored the fundamental trade-off between maximizing sensitivity to rare laundering events and containing false alarms, both of which are critical for operational deployment in financial institutions. It was observed that naïve accuracy is not decision-useful under such conditions, and that multiple complementary metrics are required to characterize performance. In particular, Balanced Accuracy, ROC-AUC, Precision, Recall, and F1 were found to convey distinct aspects of detection quality and should be interpreted jointly when setting operating targets. Accordingly, performance differences observed across models should be interpreted as reflections of distinct cost profiles rather than absolute superiority, reinforcing the need to jointly consider ranking quality, alert volume, and institutional capacity when selecting AML detection models.

RF, despite near-perfect overall accuracy, did not achieve meaningful sensitivity, indicating the limitations of tree ensembles under the present representation and the challenges posed by extreme prevalence gaps. LR attained the highest Balanced Accuracy by combining very strong recall with reasonable specificity; however, this result was accompanied by a substantial false-positive burden at the default threshold, implying the need for calibration before operationalization. XGBoost delivered moderate performance, improving recall over RF while remaining weaker than the strongest alternatives in ranking capacity. CatBoost achieved the strongest ROC-AUC and the highest precision, providing superior ranking quality and conservative positive predictions that reduced false alarms, with the trade-off of limited recall at the default operating point. Collectively, these findings indicate that no single model is universally dominant; method selection should be governed by institutional priorities, alert handling capacity, and explicit cost asymmetries. When maximizing coverage of illicit transactions is paramount, LR is the most effective baseline; when operational efficiency and false-positive control are critical, CatBoost is the most suitable candidate.

The practical implication is that model choice and threshold selection must be treated as coupled design variables within a cost-aware decision framework. Ranking-based triage is expected to offer a viable deployment pathway: scores can be used to prioritize a daily investigation budget that satisfies precision targets while preserving sufficient recall for policy or regulatory requirements. In addition, categorical information was found to be decisive for precision and ranking quality, suggesting that data enrichment through customer and account profiles, as well as network-level context, will be beneficial in production settings. In this study, all models were evaluated using a fixed default decision threshold. While explicit threshold calibration can further align model behavior with institution-specific cost constraints, such calibration is beyond the scope of the present work and is left for future research.

Limitations of the present study should be acknowledged. The evaluation relied on a synthetic dataset and did not incorporate temporal feature engineering or graph-structured signals; thus, the external validity to institution-specific environments remains to be established. Furthermore, workload and investigator behavior were not modeled; downstream impacts on case resolution and time-to-disposition could alter preferred operating points. Finally, potential issues related to data drift, class-prior shift, and alert habituation were not measured and warrant longitudinal monitoring.

Future work should expand in four main directions. First, graph-based modeling and imbalance-aware learning strategies are expected to capture structural and prevalence-related complexities beyond individual transactions. Second, temporal and behavioral features derived from transaction timestamps, on larger or real-world datasets, may improve generalization and stability. Third, extending the model portfolio to include more recent gradient-boosted alternatives and deep learning architectures would allow a broader assessment of the performance ceiling on this benchmark. In addition, deployment-oriented considerations such as threshold calibration, explainability, drift monitoring, and governance mechanisms will be essential to ensure operational feasibility and regulatory compliance.

Overall, the results indicate that machine learning can surface complex AML patterns within massive transaction streams, provided that models are calibrated to institutional objectives and embedded in workflows that reconcile investigative capacity with regulatory expectations. Under such a framework, CatBoost emerges as a compelling scorer for precision-first triage, while LR constitutes a strong high-recall baseline for coverage-first objectives; the preferred solution is therefore expected to be policy-dependent, rather than model-exclusive.

Author Contribution

The author contributed to all stages of the study.

Conflict of Interest

The author has declared no conflict of interest.

REFERENCES

[1] UNODC. "Money Laundering." United Nations Office on Drugs and Crime. https://www.unodc.org/unodc/en/money-laundering/overview.html (accessed.

[2] A. Venčkauskas, G. Š, L. Pocius, R. Brūzgienė, and A. Romanovs, "Machine Learning in Money Laundering Detection Over Blockchain Technology," IEEE Access, vol. 13, pp. 7555–7573, 2025.

[3] C.R. Alexandre and J. Balsa, "Incorporating machine learning and a risk-based strategy in an anti-money laundering multiagent system," Expert Systems with Applications, vol. 217, p. 119500, 2023/05/01/ 2023.

[4] S.S. Doddamani, K.G.K, and B. Bhowmik, "Money Laundering Detection in Imbalanced E-wallet Transactions with Threshold Optimization," in 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), 5–7 April 2024 2024, pp. 1–6.

[5] P. Sharma, A.S. Prakash, and A. Malhotra, "Application of Advanced AI Algorithms for Fintech Crime Detection," in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 24–28 June 2024 2024, pp. 1–6,.

[6] G. Konstantinidis and A. Gegov, "Deep Neural Networks for Anti Money Laundering Using Explainable Artificial Intelligence," in 2024 IEEE 12th International Conference on Intelligent Systems (IS), 29–31 Aug. 2024 2024, pp. 1–6.

[7] D.V. Kute, B. Pradhan, N. Shukla, and A. Alamri, "Explainable deep learning model for predicting money laundering transactions," International Journal on Smart Sensing and Intelligent Systems, vol. 17, no. 1, 2024, doi: 10.2478/ijssis-2024-0027.

[8] S. Ramadhan, "Harnessing machine learning for money laundering detection: a criminological theory-centric approach," Journal of Money Laundering Control, vol. 28, no. 1, pp. 184–201, 2024, doi: 10.1108/jmlc-04-2024-0083.

[9] E. Altman, J. Blanuša, L. Von Niederhäusern, B. Egressy, A. Anghel, and K. Atasu, "Realistic synthetic financial transactions for anti-money laundering models," Advances in Neural Information Processing Systems, vol. 36, pp. 29851–29874, 2023.

[10] P. Salahi, R. Shady, I. A. Doush, and M. Kandil, "Performance Analysis of Machine Learning Techniques for Detecting Money Laundering in Bitcoin Transactions," in 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), 17–19 Nov. 2024 2024, pp. 156–163.

[11] F. Wan and P. Li, "A Novel Money Laundering Prediction Model Based on a Dynamic Graph Convolutional Neural Network and Long Short-Term Memory," Symmetry, vol. 16, no. 3, p. 378, 2024. [Online]. Available: https://www.mdpi.com/2073-8994/16/3/378.

[12] S.E and S.S.S, "Identifying illicit transactions in Bitcoin Tumbler services using Supervised Machine Learning Algorithms," in 2023 12th International Conference on Advanced Computing (ICoAC), 17–19 Aug. 2023 2023, pp. 1–8.

[13] F. Irshad, T. Alkhalifah, F. Alturise, and Y. D. Khan, "GCF-MLD: Integrated Approach for Money Laundering Detection Using Machine Learning and Graph Network Analysis," IEEE Access, vol. 12, pp. 183961–183972, 2024.

[14] T. H. Phyu and S. Uttama, "Improving Classification Performance of Money Laundering Transactions Using Typological Features," in 2023 7th International Conference on Information Technology (InCIT), 16–17 Nov. 2023 2023, pp. 520–525.

[15] H. Gandhi, K. Tandon, S. Gite, B. Pradhan, and A. Alamri, "Navigating the Complexity of Money Laundering: Anti–money Laundering Advancements with AI/ML Insights," International Journal on Smart Sensing and Intelligent Systems, vol. 17, no. 1, 2024.

[16] M.E. Lokanan, "Predicting money laundering sanctions using machine learning algorithms and artificial neural networks," Applied Economics Letters, vol. 31, no. 12, pp. 1112–1118, 2024/07/11 2024.

[17] E.R. Altman. IBM Transactions for Anti Money Laundering (AML): LI-Small_Trans.csv, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/ealtman2019/ibm-transactions-for-anti-money-laundering-aml

[18] K.P. Fourkiotis and A. Tsadiras, "Future Internet Applications in Healthcare: Big Data-Driven Fraud Detection with Machine Learning," Future Internet, vol. 17, no. 10, p. 460, 2025.

[19] A. Magklaras, C. Gogos, P. Alefragis, and A. Birbas, "Enhancing Parameters Tuning of Overlay Models with Ridge Regression: Addressing Multicollinearity in High-Dimensional Data," Mathematics, vol. 12, no. 20, p. 3179, 2024.

[20] W. Sun, Q. Shen, Y. Gao, Q. Mao, T. Qi, and S. Xu, "Objective over Architecture: Fraud Detection Under Extreme Imbalance in Bank Account Opening," Computation, vol. 13, no. 12, p. 290, 2025.

[21] H.-S. Jang, "Ensemble Learning Model for Industrial Policy Classification Using Automated Hyperparameter Optimization," Electronics, vol. 14, no. 20, p. 3974, 2025.

[22] J. He, Z. Li, and L. Yin, "An Efficient and Accurate Random Forest Node-Splitting Algorithm Based on Dynamic Bayesian Methods," Machine Learning and Knowledge Extraction, vol. 7, no. 3, p. 70, 2025.

[23] F. Imam, P. Musilek, and M. Z. Reformat, "Parametric and Nonparametric Machine Learning Techniques for Increasing Power System Reliability: A Review," Information, vol. 15, no. 1, p. 37, 2024.

[24] M.B. Ayanoglu and I. Uysal, "PWFS: Probability-Weighted Feature Selection," Electronics, vol. 14, no. 11, p. 2264, 2025.

[25] R. García-Carretero, R. Holgado-Cuadrado, and Ó. Barquero-Pérez, "Assessment of Classification Models and Relevant Features on Nonalcoholic Steatohepatitis Using Random Forest," Entropy, vol. 23, no. 6, p. 763, 2021.

[26] M. Sasi, O.R. Adegboye, and A. Alzubi, "Explainable and Optimized Random Forest for Anomaly Detection in IoT Networks Using the RIME Metaheuristic," Electronics, vol. 14, no. 22, p. 4465, 2025.

[27] A. Kundu, S.G. Kundu, S.K. Sahu, and N.D. Badgayan, "Leveraging Azure Automated Machine Learning and CatBoost Gradient Boosting Algorithm for Service Quality Prediction in Hospitality," Computers, vol. 14, no. 2, p. 32, 2025.

[28] G. Marín Díaz, "A Unified Fuzzy–Explainable AI Framework (FAS-XAI) for Customer Service Value Prediction and Strategic Decision-Making," AI, vol. 7, no. 1, p. 3, 2026.

[29] E. Pellegrino et al., "Extreme Gradient Boosting Tuned with Metaheuristic Algorithms for Predicting Myeloid NGS Onco-Somatic Variant Pathogenicity," Bioengineering, vol. 10, no. 7, p. 753, 2023.

[30] A. Rashidi Nasab and H. Elzarka, "Optimizing Machine Learning Algorithms for Improving Prediction of Bridge

Deck Deterioration: A Case Study of Ohio Bridges," Buildings, vol. 13, no. 6, p. 1517, 2023.

[31] A. Andrews, N. Islam, G. Bcharah, H. Bcharah, and M. Pangasa, "Development and Internal Validation of Machine Learning Algorithms to Predict 30-Day Readmission in Patients Undergoing a C-Section: A Nation-Wide Analysis," Journal of Personalized Medicine, vol. 15, no. 10, p. 476, 2025.

[32] F. Milella, L. Famiglini, G. Banfi, and F. Cabitza, "Application of Machine Learning to Improve Appropriateness of Treatment in an Orthopaedic Setting of Personalized Medicine," Journal of Personalized Medicine, vol. 12, no. 10, p. 1706, 2022.

[33] Z. Chen and W. Fan, "A Freeway Travel Time Prediction Method Based on an XGBoost Model," Sustainability, vol. 13, no. 15, p. 8577, 2021.

[34] S.S. Abinayaa et al., "Securing the Edge: CatBoost Classifier Optimized by the Lyrebird Algorithm to Detect Denial of Service Attacks in Internet of Things-Based Wireless Sensor Networks," Future Internet, vol. 16, no. 10, p. 381, 2024.

[35] P. König et al., "Boost-Classifier-Driven Fault Prediction Across Heterogeneous Open-Source Repositories," Big Data and Cognitive Computing, vol. 9, no. 7, p. 174, 2025.

[36] Q. Wang, C. Yan, Y. Zhang, Y. Xu, X. Wang, and P. Cui, "Numerical Simulation and Bayesian Optimization CatBoost Prediction Method for Characteristic Parameters of Veneer Roller Pressing and Defibering," Forests, vol. 15, no. 12, p. 2173, 2024.

[37] M. Kopyt, P. Piotrowski, and D. Baczyński, "Short-Term Energy Generation Forecasts at a Wind Farm—A Multi-Variant Comparison of the Effectiveness and Performance of Various Gradient-Boosted Decision Tree Models," Energies, vol. 17, no. 23, p. 6194, 2024.

[38] Y. Liu, T. Yang, L. Tian, B. Huang, J. Yang, and Z. Zeng, "Ada-XG-CatBoost: A Combined Forecasting Model for Gross Ecosystem Product (GEP) Prediction," Sustainability, vol. 16, no. 16, p. 7203, 2024.

[39] Y. He, B. Yang, and C. Chu, "GA-CatBoost-Weight Algorithm for Predicting Casualties in Terrorist Attacks: Addressing Data Imbalance and Enhancing Performance," Mathematics, vol. 12, no. 6, p. 818, 2024.

[40] A. Ali et al., "Towards improved fake news detection using a hybrid RoBERTa and metadata enhanced XGBoost model," Scientific Reports, 2025/12/04 2025.

[41] R. Zhou, J. Yang, H. Liu, C. Xu, Y. Pu, and I. S. Babichuk, "A lightweight model and agent framework for fast and accurate surface defect detection in MoS2 films," Materials Science in Semiconductor Processing, vol. 193, p. 109494, 2025/07/01/ 2025.

[42] V. Sinap, "Bankruptcy Prediction with Optuna-Enhanced Ensemble Machine Learning Methods: A Comparison of Oversampling and Undersampling Techniques," (in en), Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, vol. 16, no. 1, pp. 97–113, March 2025.

[43] M.F.I. Amin, Y. Watanobe, M.M. Rahman, and A. Shirafuji, "Source Code Error Understanding Using BERT for Multi-Label Classification," IEEE Access, vol. 13, pp. 3802–3822, 2025.

[44] M. Meral, F. Ozbilgin, and F. Durmus, "Fine-Tuned Machine Learning Classifiers for Diagnosing Parkinson's Disease Using Vocal Characteristics: A Comparative Analysis," Diagnostics, vol. 15, no. 5, p. 645, 2025.

[45] M. Ait Omar, I. Etebaai, M. Taher, and A. Tawfik, "Landslide susceptibility mapping in the Bokoya Massif, Northern Morocco: A geospatial and multi-factor analysis using the analytic hierarchy process (AHP)," Scientific African, vol. 30, p. e02980, 2025/12/01/ 2025.

[46] M.H. Moharam, K. Ashraf, H. Alaa, M. Ahmed, and H. A. El-Hakim, "Real-time detection of Wi-Fi attacks using hybrid deep learning models on NodeMCU," Scientific Reports, vol. 15, no. 1, p. 32544, 2025/09/15 2025.

[47] T.D. Waele et al., "Time-Series-Based Feature Selection and Clustering for Equine Activity Recognition Using Accelerometers," IEEE Sensors Journal, vol. 23, no. 11, pp. 11855–11868, 2023.