Optimization of RNN and Tree-Based Models with Imbalance Handling for Fraud Detection in Digital Banking Transactions

Rizki Ahmad Darmawan[1], Ahmad Musyafa[2*], Murni Handayani[3]
Graduate Program of Informatics Engineering, Universitas Pamulang
Email: rizkiahmad97@gmail.com[1], dosen00668@unpam.ac.id[2], dosen02710@unpam.ac.id[3]

This study focuses on addressing the growing challenge of fraud detection in digital banking transactions, which has intensified alongside the rapid expansion of digital financial services. Fraud detection is particularly complex due to the highly imbalanced nature of transaction data, large data volumes, and intricate transaction patterns that make fraudulent activities difficult to identify accurately. Although previous research has applied a wide range of methods, from conventional machine learning techniques to advanced deep learning models, many approaches still face limitations in balancing high detection accuracy with computational efficiency. The main objective of this research is to compare the performance of Recurrent Neural Network (RNN)–based models, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bidirectional LSTM (BiLSTM), with tree-based ensemble models such as XGBoost and LightGBM in detecting fraudulent banking transactions. To enhance model effectiveness, the study implements a comprehensive data preprocessing framework that includes data cleaning, feature engineering, and techniques for handling class imbalance, particularly the use of Synthetic Minority Over-sampling Technique (SMOTE). Furthermore, model performance is optimized through systematic hyperparameter tuning using Optuna, Hyperopt, and Keras Tuner. Evaluation is conducted using metrics suitable for imbalanced datasets, such as precision, recall, F1-score, and AUC-ROC. The expected outcome is the identification of a robust and efficient fraud detection model that improves detection accuracy and sensitivity, while offering valuable insights for both academic research and practical banking applications.

Keywords: Fraud Detection, RNN, LSTM, GRU, BiLSTM, XGBoost, LightGBM, Hyperparameter Tuning, Imbalanced Data.

Corresponding Author:
Ahmad Musyafa
Graduate Program of Informatics Engineering, Universitas Pamulang
dosen00668@unpam.ac.id

1. Introduction

The rapid development of digital technology has driven significant transformations across various aspects of life, particularly within the banking industry. Digital services such as mobile banking, internet banking, and other electronic transaction systems have expanded rapidly, enabling customers to transfer funds, make cashless payments, open new accounts via smartphones, and even apply for loans without visiting bank branches. These innovations have made financial processes faster, more efficient, and more practical. However, such convenience is accompanied by increasing risks of digital crime, especially fraud, which not only causes losses to individual users but also poses broader threats to corporate stability and the sustainability of the digital banking industry.

The banking sector plays a vital role in supporting economic growth and maintaining financial system stability. Along with technological advancement, the industry has undergone significant transformation through the adoption of digital platforms. Consequently, understanding the challenges and opportunities of digital banking has become essential for ensuring long-term sustainability and competitiveness in the financial sector.

According to the Financial Services Authority (Otoritas Jasa Keuangan/OJK), digital banking services refer to banking activities conducted through electronic facilities owned by banks or digital media used by prospective and existing customers, allowing transactions to be performed independently anytime and anywhere without visiting physical branches[1]. To regulate this transformation, OJK issued Regulation No. 12/POJK.03/2021 concerning Commercial Banks, which refines Regulation No. 12/POJK.03/2018 on the Provision of Digital Banking Services. These regulations define digital banks as Indonesian legal entities that operate primarily through electronic channels with no or very limited physical offices, reflecting OJK's effort to promote innovation while ensuring sound governance.

The growth of digital banks in Indonesia has been substantial, as evidenced by institutions such as PT Bank Digital BCA, BRI Agro, PT Bank Jago, and PT Bank Seabank Indonesia. Despite the efficiency offered, digital banking systems remain vulnerable to both unintentional errors and intentional misconduct such as fraud and cybercrime. OJK Regulation No. 39 of 2019 defines fraud as deliberate actions intended to obtain unlawful benefits and cause losses to other parties. Similarly, the Association of Certified Fraud Examiners defines fraud as dishonest and illegal acts involving manipulation, concealment, or abuse of trust[2].

OJK reported a 38 percent increase in cybercrime cases in 2022 compared to the previous year, highlighting the urgency of developing robust and responsive fraud detection systems[1]. One of the primary technical challenges in fraud detection is the highly imbalanced nature of transaction data, where fraudulent transactions often represent less than one percent of the total dataset. This imbalance causes conventional classification models to favor normal transactions and overlook fraud cases, leading to high false-negative rates [3] [4].

Various strategies have been proposed to address data imbalance, including oversampling techniques such as SMOTE, undersampling, cost-sensitive learning, and ensemble methods, all of which have shown effectiveness in improving minority-class detection[3][4][5]. In addition, generative approaches such as Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) have been applied to generate synthetic fraud data under extreme imbalance conditions.

From an algorithmic perspective, two dominant approaches are tree-based models and sequential models based on Recurrent Neural Networks (RNN). Tree-based algorithms such as XGBoost and LightGBM are widely used due to their strong performance on large-scale tabular data. XGBoost demonstrates stable performance across various datasets[6][7][8], while LightGBM achieves high efficiency through Gradient-based One-Side Sampling and Exclusive Feature Bundling[9][10]. However, both models still require specialized handling of imbalanced data.

Meanwhile, sequential models such as RNN, LSTM, GRU, and BiLSTM are effective in capturing temporal dependencies between transactions, making them suitable for detecting time-based anomalous behavior. Nevertheless, improper batch sampling can bias these models toward the majority class and reduce fraud detection performance[11]. Previous studies have shown that integrating attention mechanisms or hybrid architectures can further improve sensitivity and recall in fraud detection tasks.

Despite the strong potential of these approaches, comprehensive comparisons between RNN-based and tree-based models within a unified experimental framework remain limited, particularly those integrating imbalance-handling strategies and modern hyperparameter optimization. Moreover, studies using Indonesian transaction data remain scarce, while issues such as model interpretability and distribution shift are often insufficiently addressed[5].

Therefore, this study focuses on optimizing and comparing RNN models (LSTM, GRU, BiLSTM) and tree-based models (XGBoost, LightGBM) for fraud detection in digital banking transactions. This research

integrates imbalance-handling techniques and advanced hyperparameter optimization methods such as Hyperopt and Optuna [12][13]. The ultimate objective is to provide an efficient, accurate, and practically applicable fraud detection model for the Indonesian banking industry, balancing precision, recall, computational cost, and interpretability.

2. Literature Review and Problem Statement

Literature Review

Fraud detection in digital banking transactions remains a critical research topic due to the complex characteristics of financial transaction data, including extreme class imbalance, evolving fraud patterns, concept drift, and delayed labeling. Dal Pozzolo, Caelen, and Bontempi[4] emphasize that fraud detection is fundamentally different from conventional classification problems because fraudulent events are rare and continuously adapt to defensive mechanisms. Without proper temporal validation and imbalance-aware modeling, detection systems are prone to data leakage and unrealistic performance estimation.

One of the earliest and most influential approaches to addressing fraud detection challenges is cost-sensitive learning. Bahnsen et al[3] argue that accuracy alone is insufficient in fraud detection, as the financial cost of misclassifying fraudulent transactions as legitimate (false negatives) is significantly higher than false positives. Their findings demonstrate that incorporating cost matrices into classification models leads to more realistic and economically effective fraud detection systems, particularly in highly imbalanced environments.

Tree-based models such as XGBoost and LightGBM have been widely adopted due to their robustness on large-scale tabular data. Chen[6] show that XGBoost achieves strong generalization performance through gradient boosting optimization, while Ke et al[14] highlight LightGBM's efficiency via Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). However, subsequent studies reveal that both models suffer performance degradation under extreme imbalance if no special handling is applied.

Recent research focuses on improving tree-based models through imbalance-aware optimization. Mehdari et al[15] demonstrate that adaptive hyperparameter tuning significantly enhances XGBoost's recall and robustness. Similarly, Rezaei et al.[16] introduce cost-balanced and oversampling-based loss functions for LightGBM, achieving notable improvements in recall and AUC under ultra-imbalanced conditions. These findings confirm that imbalance handling and hyperparameter optimization are critical for tree-based fraud detection models.

In parallel, deep learning approaches particularly Recurrent Neural Networks (RNN) have gained attention for their ability to model sequential transaction behavior. Fiore et al[17] show that LSTM and GRU models combined with oversampling techniques such as SMOTE can significantly improve fraud recall. However, Reddy [18] caution that naïve batch sampling may cause RNNs to predict all transactions as non-fraud, necessitating sequence-aware sampling strategies.

To further enhance sequential modeling, recent studies integrate attention mechanisms and hybrid architectures. Models such as BiLSTM-Attention, CNN–BiLSTM–Attention, and BiLSTM–Transformer consistently outperform standalone RNNs by improving recall, AUC, and interpretability[19][20][21]. These architectures allow models to focus on critical temporal and feature-level patterns associated with fraudulent behavior.

Beyond RNN and tree-based methods, hybrid and generative approaches have emerged. Carcillo et al.[5] propose a hybrid unsupervised–supervised framework combining autoencoders and LightGBM, demonstrating superior adaptability to evolving fraud patterns. More advanced models such as Regularised

Memory Graph Attention Networks and Balanced Variational Autoencoders further highlight the importance of representation learning and relational modeling in fraud detection[22].

Despite these advancements, existing studies often focus on isolated model types or single imbalance-handling strategies. Comprehensive comparisons between RNN-based and tree-based models within a unified experimental framework incorporating imbalance handling, hyperparameter optimization, and realistic evaluation metrics remain limited, particularly in the context of Indonesian digital banking transactions. This gap underscores the necessity for integrative research that systematically evaluates both approaches to identify models that are not only accurate but also computationally efficient and practically deployable in real-world banking systems.

Problem Statement

The rapid growth of digital banking transactions in Indonesia has significantly increased exposure to fraudulent activities, creating substantial financial losses and undermining customer trust in digital financial services. Despite regulatory frameworks and technological advancements, fraud detection in digital banking environments remains a persistent challenge due to the complex, large-scale, and highly imbalanced nature of transaction data, where fraudulent transactions represent only a very small fraction of overall observations.

Existing fraud detection systems frequently suffer from high false negative rates, primarily caused by extreme class imbalance that biases predictive models toward non-fraud transactions. Traditional classification approaches tend to prioritize overall accuracy, which is insufficient in fraud detection contexts where the cost of misclassifying fraudulent transactions is substantially higher than false positives. Although imbalance-handling techniques such as SMOTE and cost-sensitive learning have been proposed, their effectiveness varies depending on model architecture and data characteristics.

Moreover, prior studies predominantly examine fraud detection models in isolation. Tree-based models such as XGBoost and LightGBM are recognized for their efficiency and strong performance on tabular data, yet their robustness decreases under extreme imbalance without tailored optimization. Conversely, sequential deep learning models based on Recurrent Neural Networks (LSTM, GRU, and BiLSTM) demonstrate superior capability in capturing temporal transaction patterns but are prone to bias and instability when trained on imbalanced datasets without appropriate sampling strategies.

Another critical limitation in existing research lies in the lack of comprehensive comparative studies that evaluate RNN-based and tree-based models within a unified experimental framework. Most prior works do not simultaneously integrate imbalance-handling strategies and modern hyperparameter optimization techniques, nor do they sufficiently address methodological issues such as temporal validation, data leakage, and model interpretability. Additionally, empirical studies utilizing real-world or localized Indonesian banking transaction data remain limited, reducing the practical applicability of existing findings.

Consequently, there is currently no fraud detection model that can be considered simultaneously accurate, robust, computationally efficient, and practically deployable for digital banking transactions in Indonesia. This condition highlights the urgent need for a comprehensive study that systematically compares and optimizes RNN-based and tree-based models by integrating effective imbalance-handling techniques and advanced hyperparameter tuning. Addressing this gap is essential to develop a fraud detection system capable of minimizing false negatives, improving detection sensitivity, and supporting secure and sustainable digital banking operations.

3. Method

Requirements Analysis

The requirements analysis aims to identify all essential components needed to conduct this research systematically, including data characteristics, data collection methods, data sources, preprocessing techniques, and algorithm development strategies. This stage ensures that the research process is well-structured and aligned with the study objectives, particularly in developing an effective fraud detection model for digital banking transactions.

Population and Sample

The population of this study consists of all digital banking transactions conducted through electronic channels such as mobile banking, internet banking, and digital payment platforms. Given the massive and dynamic nature of transaction data, this study employs a purposive sampling approach using online debit transaction data obtained from one digital bank in Indonesia.

The selected dataset reflects a highly imbalanced distribution, where non-fraud transactions significantly outnumber fraud transactions, thereby representing real-world fraud detection challenges faced by the banking industry[4][3]. The sample selection considers data completeness, availability of fraud labels, and the ability to capture comprehensive transactional behavior patterns.

Data Collection Method

Data were collected using secondary documentation and observational approaches. The dataset consists of structured historical records of online debit transactions, including transaction timestamps, transaction amounts, merchant information, transaction channels, and fraud labels. All data were anonymized to ensure customer privacy and compliance with banking data protection standards. Preprocessing steps included data cleaning, handling missing values, normalization, and time-format transformation. The use of secondary transactional data is justified as it accurately represents real operational conditions in digital banking fraud detection[1].

Data Type and Source

This study uses quantitative secondary data derived from historical online debit transactions recorded in the internal information system of a digital bank in Indonesia. Key variables used in this research are summarized in Table 1.

Table 1. Transaction Variables

No	Variable Name	Description	Role
1	TRX_DATE	Transaction date	Temporal feature
2	TRX_TIME	Transaction time	Temporal feature
3	CHANNEL	Transaction channel	Channel identification
4	TERMINAL_CODE	Terminal identifier	Location proxy
5	AMOUNT	Transaction value	Core numeric feature
6	ISSUER	Issuing bank	Institutional feature
7	ACQUIRER	Acquiring bank	Institutional feature
8	INFO_MERCHANT	Merchant information	Merchant profiling
9	MCC	Merchant Category Code	Business category
10	SETTLEMENT_DATE	Settlement date	Temporal feature

Initial Data Processing

Initial processing involved constructing derived behavioral indicators (A1–A6) to capture suspicious transaction patterns based on transaction frequency, amount, and timing. These indicators reflect common fraud scenarios such as unusually large transactions, repeated transactions within short time windows, and abnormal transaction frequencies, consistent with fraud behavior patterns reported by ACFE and prior studies.

Table 2. Suspicious Transaction Behavior Variables

Variable	Description
A1	1 transaction exceeding IDR 5,000,000
A2	3 transactions exceeding IDR 2,000,000 between 00:00–06:00
A3	Accumulated transactions totaling IDR 10,000,000 within 1 day
A4	5 repeated transactions with identical amounts
A5	100 transactions within 1 hour
A6	20 transactions within 1 day

These features were transformed into numerical and categorical representations during feature engineering to enhance model discrimination capability between fraud and non-fraud transactions.

Algorithm Development

This study compares two main modeling approaches: Recurrent Neural Networks (LSTM, GRU, BiLSTM) and Tree-Based Models (XGBoost, LightGBM). The dataset was split into training and testing sets using an 80:20 ratio with time-based splitting to prevent data leakage[4]. Imbalance handling techniques, including SMOTE and cost-sensitive learning, were applied to mitigate class bias[3]. Hyperparameter optimization was conducted using Bayesian Optimization via Optuna to improve model generalization[12]. Model performance was evaluated using precision, recall, F1-score, and ROC-AUC metrics.

Model Evaluation and Discussion

Model evaluation was conducted on unseen test data to ensure objective performance assessment. Results indicate that imbalance-handling techniques significantly improve fraud detection sensitivity. RNN-based models, particularly BiLSTM, demonstrate strong performance in capturing temporal fraud patterns, whereas tree-based models offer superior computational efficiency and interpretability. Overall, optimized tree-based models provide the best balance between accuracy, computational cost, and explainability, while RNN models excel in detecting sequential fraud behavior. These findings support previous research emphasizing the importance of imbalance handling and hyperparameter optimization in fraud detection systems.

4. Results and Discussion

Research Dataset

The study used 4,999 historical digital-banking debit transactions. Because the raw transactions were not provided with official fraud labels, the study adopted a rule-based pseudo-labeling strategy to approximate transaction risk, reflecting real-world constraints where fraud labels can be delayed and require manual investigation[4]. The pseudo-label is then used as the target for supervised modeling.

Table 3. Online Debit Transaction Dataset

DATE	TIME	NO_CARD	AMOUNT	APPROVAL	MCC
240313	08:28:16	5260512001xxxxxx	331650	828164	5541
240313	19:23:27	5260513001xxxxxx	83000	192328	5814
240313	13:11:56	5260514001xxxxxx	5550000	131156	8099
240313	14:17:38	5260515001xxxxxx	1415700	141738	5813
240313	00:03:51	5260516001xxxxxx	724881	351432	5541
240313	21:27:54	5260517001xxxxxx	321930	212754	5813
240313	12:44:41	5260518001xxxxxx	200000	124441	5541
240313	15:57:20	5260519001xxxxxx	350000	155720	5541
240313	19:20:56	5260510001xxxxxx	400000	192056	4112
240313	21:05:00	5260511001xxxxxx	955185	210500	5812

The dataset in the table above displays only a few features from the original dataset. The absence of class labels or fraud and non-fraud statuses reflects the real-world conditions in the banking industry, where the process of labeling fraudulent transactions is often delayed and requires manual investigation. Therefore, this dataset is considered relevant for testing the effectiveness of anomaly pattern-based fraud detection models.

Initial Processing and Rule-Based Risk Labeling

In the initial data processing stage, this study adopts a two-stage approach to establish transaction risk assessment. The first stage applies a rule-based approach using features A1–A6 as risk indicators defined based on domain knowledge and general financial transaction policies. Each transaction is evaluated using deterministic rules to generate an initial risk label. In the second stage, a machine learning approach is implemented using additional transactional behavior features not applied in the first stage. These features are used to build a predictive model that learns transaction behavior patterns and predicts the initial risk labels, enabling adaptive risk generalization based on transaction characteristics.

For rule-based risk labeling, transaction data are sorted by card number (NO_CARD) and transaction time (TRX_DATETIME) to preserve chronological order. Features A1–A6 are then constructed as binary variables representing suspicious transaction behaviors, such as high-value transactions, high-frequency transactions, abnormal transaction times, and repeated transaction amounts within a specific period. The resulting structured dataset provides initial risk indicators to support subsequent modeling and analysis.

Table 4. Rule-Based Risk Labels

Code	Suspicious behavior indicator	Count
A1	>5 transactions within 30 minutes	34
A2	>2 transactions within <10 minutes/day	0
A3	>5 transactions within <10 minutes/day	0
A4	>3 transactions within 1 hour	16
A5	>10 transactions within 1 hour	0
A6	>20 transactions/day	0
Total		50

Table 4. presents the results of the rule-based risk labeling process using six suspicious behavior indicators (A1–A6). A transaction is classified as high risk if it meets at least one anomalous indicator. The results show that 50 transactions (1.00%) are labeled as high risk, while 4,949 transactions (99.00%) are classified as low risk, indicating a highly imbalanced dataset.

Only two indicators appear in the data: A1 (single high-value transactions) with 34 occurrences and A4 (repeated transactions with identical amounts) with 16 occurrences. All high-risk transactions satisfy only one anomaly indicator, with no overlapping risk patterns.

Table 5. Dataset Of Rule-Based Risk Label Formation Results

	amount	time_of_day	day_of_week	merchant_category	transaction_type	previous_failed_attempts
0	399300	11.816667	2	12	0	0
1	574600	9.2	2	11	0	0
2	1250000	7.4	2	2	0	0
3	99000	17.166667	2	12	0	0
4	317400	13.7	2	19	0	0

account_age_days	transaction_frequency	balance_before	balance_after	is_foreign	device_type
503	5	8.23E+05	4.24E+05	0	1
839	4	1.68E+06	1.11E+06	0	2
623	9	3.25E+06	2.00E+06	0	0
669	8	2.37E+05	1.38E+05	0	3
698	5	5.50E+05	2.33E+05	1	0

ip_address_country	velocity_1h	velocity_24h	risk_label
23	386	4999	0
32	249	4999	0
28	135	4999	0
21	386	4999	0
4	372	4999	0

Machine Learning–Based Transaction Risk Modeling

In the second stage, the dataset from the previous phase is used as input for machine learning models. The dataset consists of 4,999 transactions with 16 attributes, including one target variable (risk label) and 15 predictor features. The risk label is a pseudo-label generated from the rule-based approach in the first stage and represents an initial risk indication rather than an actual fraud label.

The class distribution is highly imbalanced, with high-risk transactions accounting for only 1.00% of the data. To prevent data leakage, all rule-based indicators (A1–A6) are excluded from the modeling process. The models are trained using remaining behavioral features related to transaction amount, frequency, velocity, account age, and temporal, device, and location information.

Numerical and behavioral features are constructed to represent transaction, card, and temporal characteristics. Categorical variables are normalized and encoded using label encoding. Additional behavioral features, including prior transaction failures, estimated foreign transactions, device type, transaction frequency, and hourly and daily transaction velocity, are incorporated.

Table 6. Final Dataset Of Machine Learning-Based Transaction Risk Modeling

	amount	time_of_day	day_of_week	merchant_category	transaction_type	previous_failed_attempts
count	5.00E+03	4999	4999	4999	4999	4999
mean	3.76E+05	14.749813	2	7.818764	0.032807	0
std	1.74E+06	4.468148	0	6.669704	0.254066	0
min	1.00E+00	0.016667	2	0	0	0
25%	7.45E+04	11.566667	2	1	0	0
50%	1.55E+05	14.916667	2	11	0	0
75%	3.17E+05	18.366667	2	12	0	0
max	1.00E+08	23.55	2	19	2	0

account_age_days	transaction_frequency	balance_before	balance_after	is_foreign
4999	4999	5.00E+03	5.00E+03	4999
531.425685	6.112623	8.57E+05	4.81E+05	0.270454
288.977445	2.417076	4.70E+06	3.01E+06	0.444239
30	1	1.55E+00	5.52E-01	0
282	4	1.60E+05	8.22E+04	0
532	6	3.48E+05	1.79E+05	0
783	8	7.25E+05	3.93E+05	1
1029	15	2.95E+08	1.95E+08	1

device_type	ip_address_country	velocity_1h	velocity_24h
4999	4999	4999	4999
1.512102	22.423085	313.318464	4999
1.115682	15.046202	84.27941	0
0	0	13	4999
1	8	257	4999
2	23	362	4999
2	36	374	4999
3	49	386	4999

The dataset and feature configuration at this stage form the basis for the implementation and evaluation of the RNN and tree-based models discussed in the following subsection.

Data Preprocessing

In this stage, the final dataset is separated into input features (X) and the target variable (y). All attributes except *risk_label* are used as input features, while *risk_label* serves as the target variable representing transaction risk. The label is a pseudo-label derived from suspicious transaction behavior patterns (A1–A6) rather than an actual fraud ground truth.

Basic statistical analysis is conducted to examine the number of samples, features, and class distribution, confirming that the dataset is imbalanced with a small proportion of high-risk transactions. Descriptive statistics, missing value checks, and data type validation are performed to ensure data consistency and readiness for machine learning modeling.

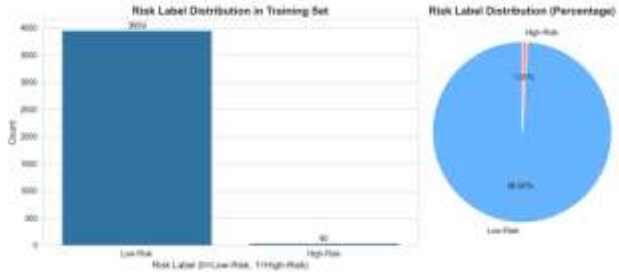

Figure 1 Distribution of Risk Label Classes

The figure illustrates the distribution of risk labels (*risk_label*) in the training data using bar and pie charts. Low-risk transactions dominate the dataset with 3,959 records, while only 40 transactions are labeled as high risk. Approximately 99% of the training data represent normal transactions, and only about 1% indicate potential fraud. This highly imbalanced class distribution is a common characteristic of fraud detection problems and requires careful modeling to prevent bias toward the majority class and ensure effective detection of high-risk transactions.

Handling Data Imbalance

1. SMOTE Method

To address data imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training dataset to enhance the representation of the minority class. SMOTE is performed exclusively on the training data to prevent data leakage. This method generates synthetic samples for the high-risk transaction class, increasing its proportion to approximately 50% relative to the normal transaction class.

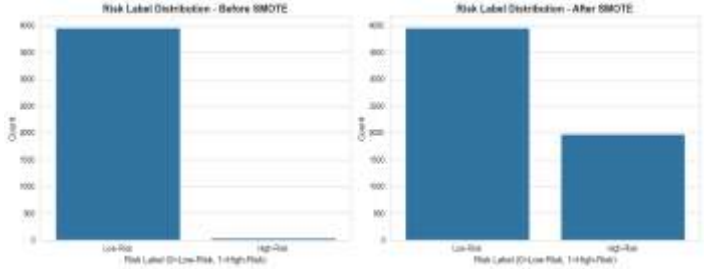

Figure 2. Risk Label Distribution after SMOTE

The figure compares the risk label distribution in the training data before and after applying the Synthetic Minority Over-sampling Technique (SMOTE). Prior to SMOTE, the training data are heavily dominated by low-risk (normal) transactions, with very few high-risk (fraud-indicative) transactions, which may bias the model toward learning normal patterns. After SMOTE, the proportion of high-risk transactions increases

substantially, resulting in a more balanced class distribution. This improvement is achieved by generating synthetic minority samples based on feature similarity rather than simple duplication, leading to a more representative training set while preserving the original distribution in the test data.

2. Cost-Sensitive Learning

As an alternative, cost-sensitive learning is applied without oversampling. The training data retain the original imbalanced distribution, and class imbalance is addressed by assigning higher misclassification costs to the minority class. Class distribution ratios are used to set the scale_pos_weight parameter in XGBoost, while automatic class weighting is applied in neural network and LightGBM models. This approach allows the models to learn from realistic data distributions while improving sensitivity to fraud transactions, enabling objective comparison with the SMOTE-based scenario during model evaluation.

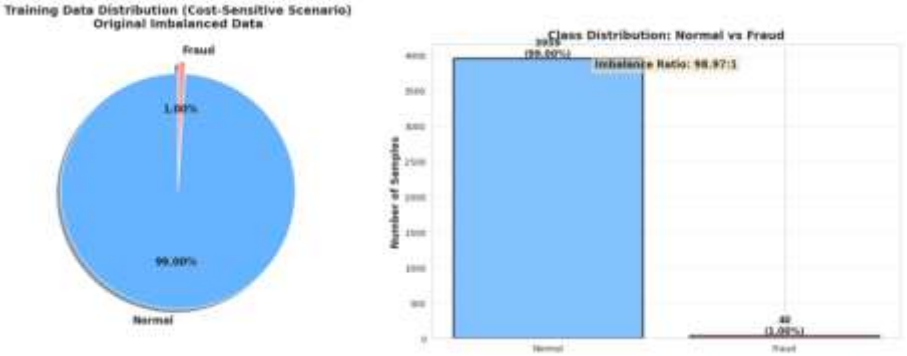

Figure 3. Risk Label Distribution after Cost-Sensitive Learning

Figure 3 illustrates the training data distribution under the cost-sensitive learning scenario using the original, imbalanced dataset. The pie chart shows that normal transactions account for 99.00% of the data, while fraud transactions represent only 1.00%, indicating that fraud is a rare event. The bar chart confirms this imbalance in absolute terms, with approximately 3,959 normal transactions and only 40 fraud transactions, resulting in an extreme imbalance ratio of 98.97:1. This visualization highlights the core challenge in fraud detection modeling and justifies the use of cost-sensitive learning, which increases penalties for misclassifying the minority (fraud) class without altering the original data distribution.

Model Implementation

This study implements two modeling approaches: Recurrent Neural Networks (RNN) and tree-based models. RNN models are trained iteratively using backpropagation to learn sequential transaction patterns, while tree-based models emphasize hyperparameter tuning to achieve optimal performance and mitigate overfitting and underfitting. This design ensures that each model is aligned with the characteristics of the transaction data and the objectives of fraud detection.

1. RNN Models

a. LSTM

The LSTM model with SMOTE achieves a ROC-AUC of 0.93, indicating good discrimination capability. However, fraud detection performance remains limited, with precision, recall, and F1-score all at 0.60. The high overall accuracy is largely influenced by the dominance of normal transactions, suggesting that SMOTE does not substantially improve minority class detection for LSTM. When cost-sensitive learning is applied, LSTM performance on the fraud class improves. The model attains a ROC-AUC of 0.986 and a fraud recall of 0.80, indicating higher sensitivity. Nevertheless, lower precision (0.53) reflects increased false positives. Overall, the cost-sensitive approach enhances fraud focus compared to SMOTE, albeit with reduced accuracy on normal transactions.

b. GRU
The GRU model with SMOTE shows strong performance, achieving a ROC-AUC of 0.9899, with fraud recall of 0.80 and an F1-score of 0.67. This indicates that GRU captures sequential patterns more effectively than unidirectional LSTM. However, some normal transactions are still misclassified as fraud. Under cost-sensitive learning, GRU achieves a fraud recall of 0.90 with a ROC-AUC of 0.986, representing the highest fraud detection sensitivity among GRU scenarios. The trade-off is lower precision (0.53), indicating an increase in false positives.

c. Bidirectional LSTM (BiLSTM)
BiLSTM with SMOTE provides the most balanced performance among RNN models. With a ROC-AUC of 0.984, the model achieves a fraud precision of 0.67, recall of 0.80, and the highest F1-score (0.73). These results demonstrate that bidirectional sequence learning improves fraud pattern representation and offers the best balance between sensitivity and precision. In the cost-sensitive setting, BiLSTM maintains a fraud recall of 0.80 with a ROC-AUC of 0.9762, but precision decreases to 0.53, resulting in an F1-score of 0.64. While fraud detection improves, further refinement is required to reduce false positives.

2. Tree-Based Models

a. XGBoost
XGBoost with SMOTE delivers near-perfect performance, achieving a ROC-AUC of 1.00 and a fraud recall of 1.00. With a precision of 0.83 and an F1-score of 0.91, this model demonstrates excellent balance between sensitivity and accuracy, outperforming all RNN-based models. Under cost-sensitive learning, XGBoost maintains strong performance with a ROC-AUC of 0.9995, fraud recall of 0.80, and precision of 0.89, resulting in an F1-score of 0.84. This indicates a stable and well-balanced detection capability without compromising performance on the normal class.

b. LightGBM
LightGBM achieves perfect classification performance under both SMOTE and cost-sensitive scenarios, with all evaluation metrics equal to 1.00. These results indicate exceptional capability in handling data imbalance. However, given the limited number of fraud samples, this perfect performance should be interpreted cautiously to ensure that overfitting does not occur.

Model Evaluation

Model evaluation is conducted to assess the effectiveness of each approach in detecting fraud under imbalanced data conditions. Based on the confusion matrices, RNN models (LSTM, GRU, and Bidirectional LSTM) consistently achieve a fraud recall of 0.80, indicating that 8 out of 10 fraud transactions are correctly detected. However, two fraud cases remain undetected (false negatives), which poses potential risk in fraud detection. In addition, RNN models exhibit relatively higher false positives, leading to lower precision and F1-scores.

Among the RNN variants, Bidirectional LSTM demonstrates comparatively better performance than unidirectional LSTM and GRU, as reflected by higher precision and F1-scores. This suggests that bidirectional sequence modeling captures transaction patterns more comprehensively. Nevertheless, the overall performance of RNN models remains inferior to tree-based models in balancing sensitivity and prediction accuracy.

In contrast, tree-based models, particularly XGBoost and LightGBM, achieve substantially superior performance across all evaluation metrics. XGBoost attains a fraud recall of 1.00, successfully identifying all fraud transactions without false negatives, while LightGBM achieves a recall of 0.90, missing only one fraud case. The near-perfect ROC-AUC values further confirm the exceptional discriminative capability of tree-based models in separating normal and fraudulent transactions.

Table 7. Comprehensive Model Comparison

COMPREHENSIVE MODEL COMPARISON TABLE

Metrics for Fraud Transaction Detection:
- Precision (Fraud): Accuracy of fraud predictions
- Recall (Fraud): Ability to detect fraud transactions (minimize false negatives)
- F1-Score: Balance between precision and recall
- ROC-AUC: Overall discriminative ability for risk classification

Model	Imbalance_Method	Hyperparameter Method	Precision Fraud	Recall Fraud	ROC_AUC
LightGBM	SMOTE+Cost-Sensitive	Default	1.00	1.0	1.00
XGBoost	SMOTE+Cost-Sensitive	GridSearchCV	0.83	1.0	0.99
XGBoost	SMOTE+Cost-Sensitive	Default	0.83	1.0	1.00
LightGBM	SMOTE+Cost-Sensitive	GridSearchCV	1.00	1.0	1.00
GRU	SMOTE+Cost-Sensitive	Default	0.57	0.8	0.98
BiLSTM	SMOTE+Cost-Sensitive	Default	0.66	0.8	0.98
LSTM	SMOTE+Cost-Sensitive	Default	0.60	0.6	0.93

The comprehensive comparison table further reinforces these findings, showing that XGBoost and LightGBM achieve the highest F1-scores among all evaluated models. Both SMOTE and cost-sensitive learning effectively improve sensitivity to the fraud class; however, their impact is substantially more pronounced when applied to tree-based models than to RNNs. In RNN models, improvements in recall are often accompanied by reduced precision, resulting in a less optimal performance balance.

Overall, this evaluation indicates that tree-based models represent the most effective and stable approach for fraud transaction detection on the analyzed dataset. While RNNs offer advantages in capturing sequential patterns, tree-based models are better suited to datasets with very limited fraud samples, as they exploit feature characteristics more efficiently. These findings highlight the importance of selecting model architectures that align with data characteristics to develop reliable and practical fraud detection systems.

1. Confusion Matrix SMOTE and Cost Sensitive Learning

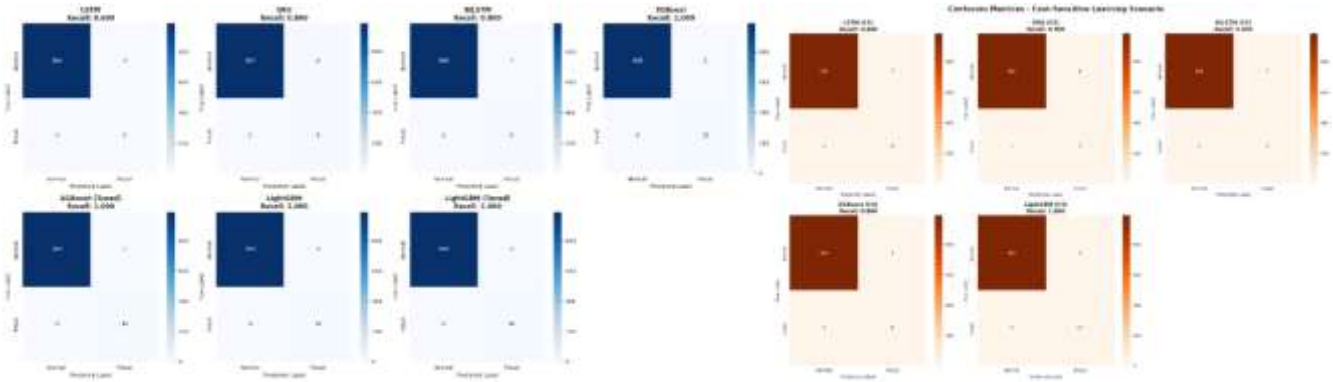

Figure 4. Confusion Matrix SMOTE and Cost Sensitive Learning

Figures 4 present the confusion matrices for the SMOTE baseline and cost-sensitive learning scenarios, with fraud recall as the primary evaluation metric. Under the SMOTE baseline, the LSTM model achieves a recall of 0.60, indicating limited sensitivity to fraud. GRU and Bidirectional LSTM improve recall to 0.80, detecting most fraud cases, with Bidirectional LSTM producing fewer false positives. In contrast, tree-based models clearly outperform RNNs, as both XGBoost and LightGBM achieve perfect fraud recall (1.00), successfully identifying all fraud transactions without false negatives.

In the cost-sensitive learning scenario, RNN models show improved sensitivity, with LSTM and Bidirectional LSTM reaching a recall of 0.80 and GRU achieving the highest recall among RNNs at 0.90. However, these gains remain inferior to tree-based performance. XGBoost attains a recall of 0.80, while LightGBM again achieves perfect fraud recall (1.00) with strong performance on normal transactions. Overall, the results confirm that tree-based models, particularly LightGBM, consistently outperform RNN models in minimizing false negatives under both SMOTE and cost-sensitive settings, highlighting their robustness for fraud detection in highly imbalanced datasets.

2. ROC Curves SMOTE and Cost-Sensitive

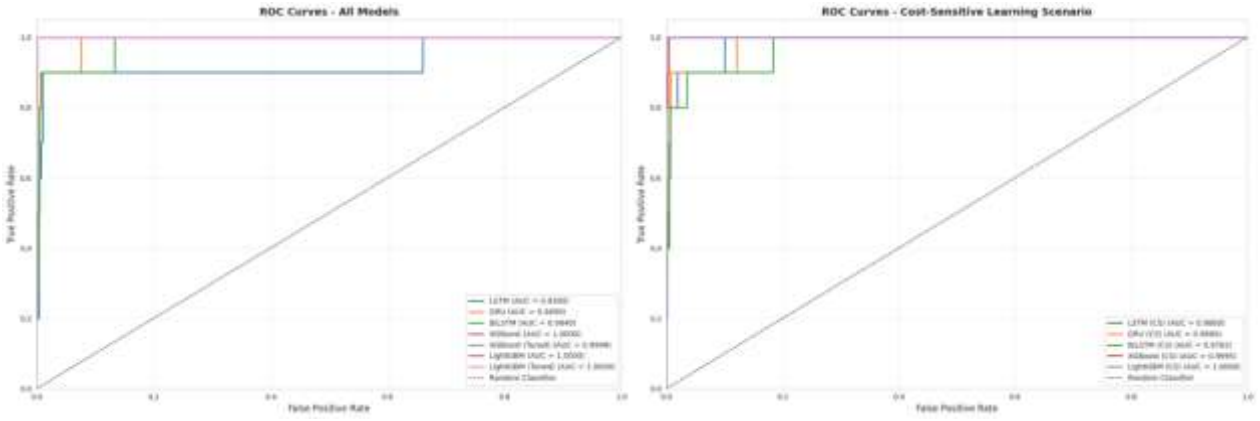

Figure 5. ROC Curves SMOTE and Cost-Sensitive

The ROC curve analysis under both SMOTE and cost-sensitive learning scenarios shows that all models demonstrate strong discriminative capability, with curves positioned well above the diagonal line. Under SMOTE, GRU and Bidirectional LSTM outperform LSTM among RNN models, indicating more effective utilization of synthetic samples. However, tree-based models consistently achieve superior performance, with XGBoost and LightGBM reaching near-perfect ROC-AUC values close to or equal to 1.00.

In the cost-sensitive learning scenario, all models exhibit further improvement in discrimination, confirming that higher misclassification costs enhance sensitivity to fraud. While RNN models achieve high and relatively balanced ROC-AUC values, their performance remains slightly below that of tree-based models. XGBoost and LightGBM again dominate, producing near-perfect ROC curves with very high true positive rates and minimal false positive rates. Overall, these results confirm that both SMOTE and cost-sensitive learning are most effective when combined with tree-based models for fraud detection.

Performance Comparison

The comparative evaluation under both SMOTE and cost-sensitive learning scenarios demonstrates clear performance differences across model architectures. Overall, all models exhibit strong discriminative capability, as reflected by high ROC-AUC values. However, tree-based models, particularly XGBoost and LightGBM, consistently achieve near-perfect or perfect ROC-AUC, indicating superior class separation compared to RNN-based models.

ISSN 2828-9463

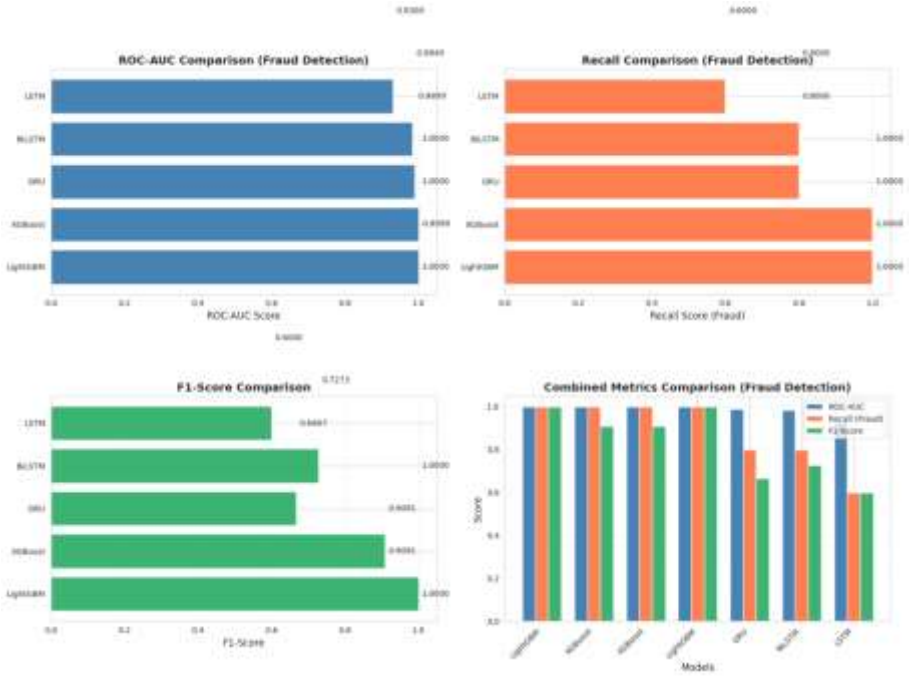

Figure 6. Performance Comparison SMOTE

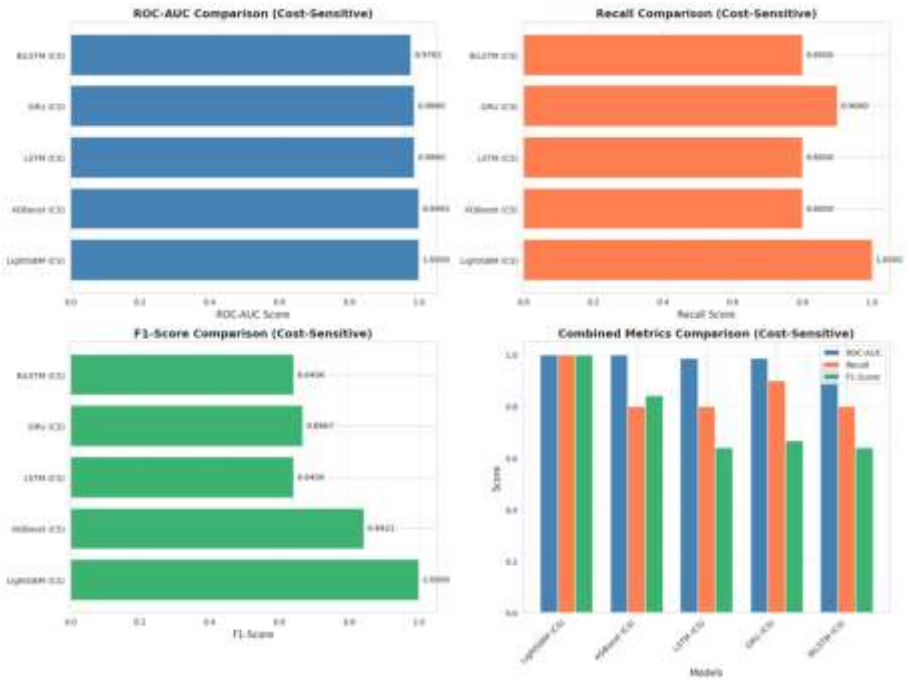

Figure 7. Performance Comparison Cost-sensitive

Under the SMOTE scenario, XGBoost and LightGBM achieve perfect fraud recall and the highest F1-scores, successfully detecting all fraud transactions without false negatives. RNN models show lower recall and F1-scores, with Bidirectional LSTM and GRU outperforming LSTM but still exhibiting higher false positive rates. Similar patterns are observed under cost-sensitive learning, where recall generally improves across models, especially for GRU. Nevertheless, LightGBM remains the most robust model, achieving perfect recall and F1-score, followed by XGBoost with strong but slightly lower balance.

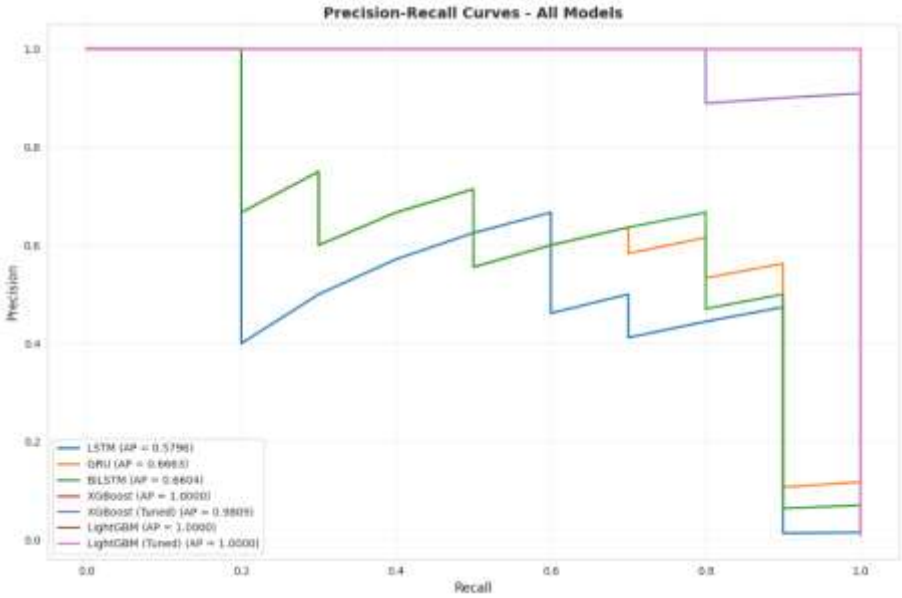

Table 8. Precision Recall Curve SMOTE

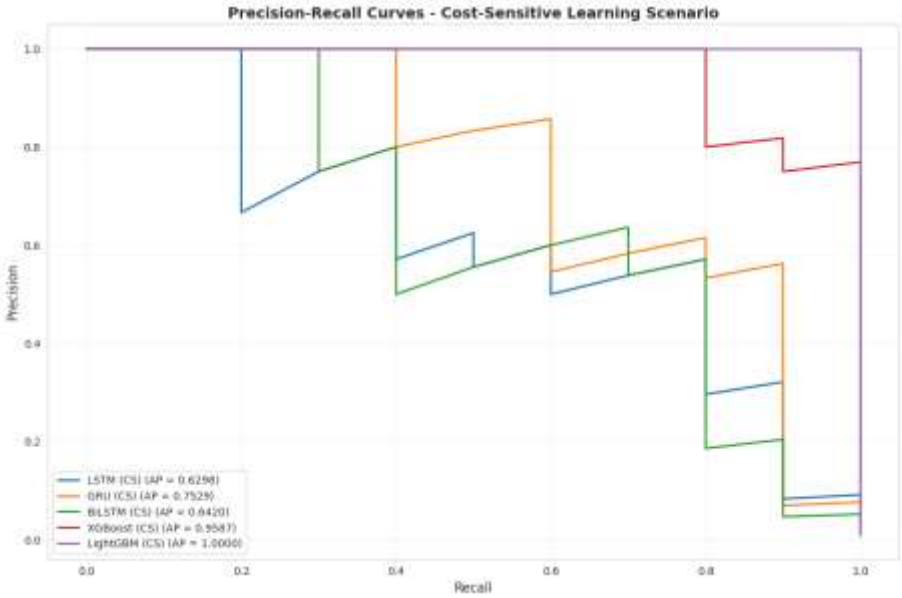

Table 9. Precision Recall Curve Cost-sensitive

Precision–recall analysis further confirms these findings. Tree-based models maintain high precision even at maximum recall, whereas RNN models experience a noticeable trade-off between sensitivity and precision, particularly under cost-sensitive learning. This indicates that improvements in fraud detection for RNNs are often accompanied by increased false positives.

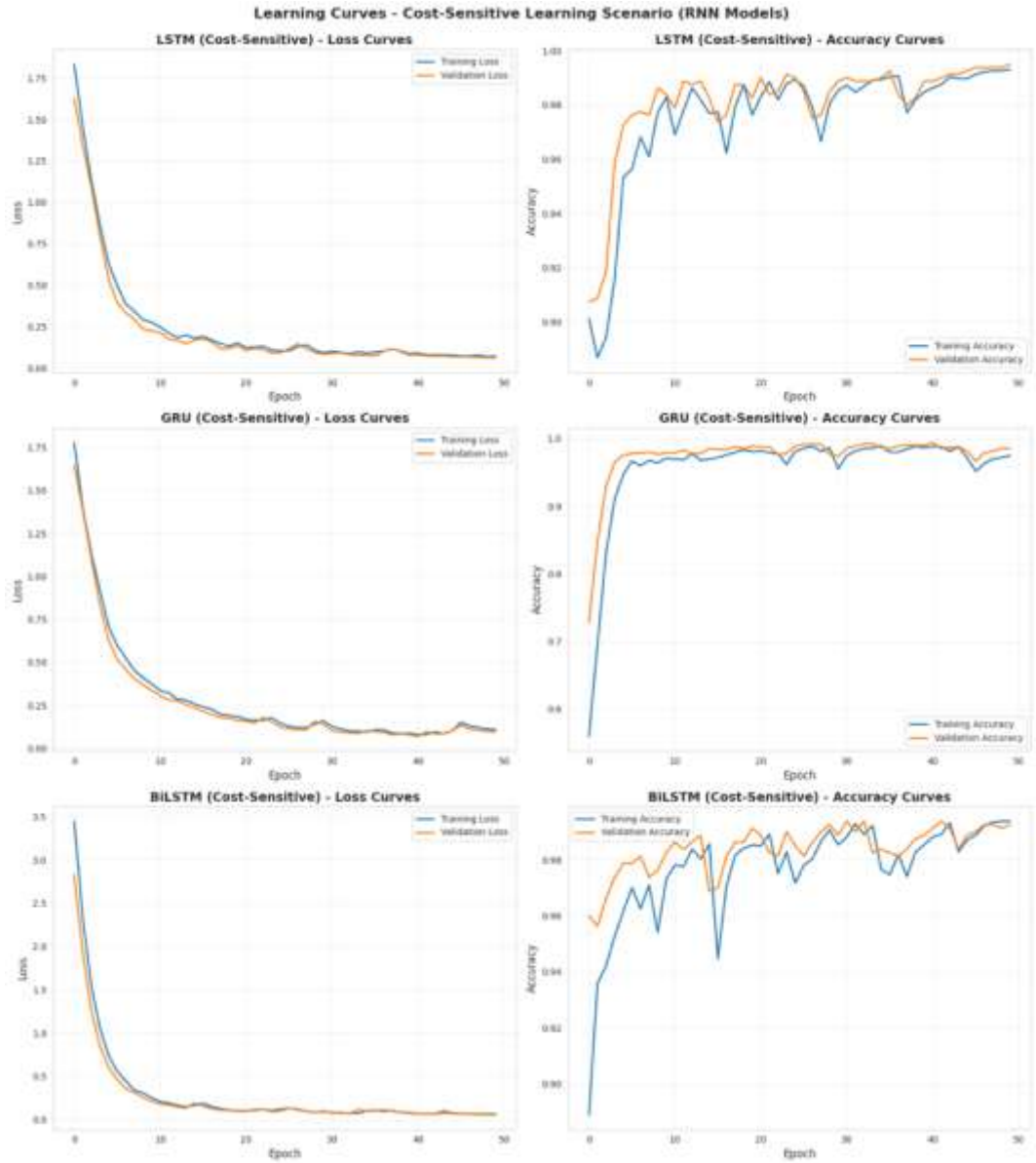

Table 10. Learning Curves Cost-sensitive

Training dynamics analysis shows that GRU converges faster and more stably than LSTM and Bidirectional LSTM, while the latter exhibits greater sensitivity due to architectural complexity. Overall, these results confirm that tree-based models, especially LightGBM, provide the most effective, stable, and balanced performance for fraud detection in highly imbalanced transaction data.

Discussion

This section elaborates on the research findings obtained from the preceding stages, encompassing data preprocessing, class imbalance handling, model development, and performance evaluation in fraud detection. The discussion highlights how these stages are interrelated and how they collectively influence the effectiveness of fraud detection systems.

Data preprocessing is a crucial preliminary step that ensures data quality prior to modeling. Procedures such as data cleaning, feature transformation, normalization, and structural alignment produced a coherent and model-ready dataset. Effective preprocessing minimizes bias caused by invalid records or heterogeneous feature scales and provides a solid foundation for achieving stable and reliable model performance.

A key challenge addressed in this study is the severe class imbalance, where fraudulent transactions represent only a small fraction of the dataset. To mitigate this issue, SMOTE and cost-sensitive learning were applied. Both methods enhance model responsiveness to fraud, though through different mechanisms. SMOTE increases minority class representation via oversampling, while cost-sensitive learning emphasizes fraud detection by assigning higher penalties to misclassification. These approaches produce distinct trade-offs between recall and precision across different models.

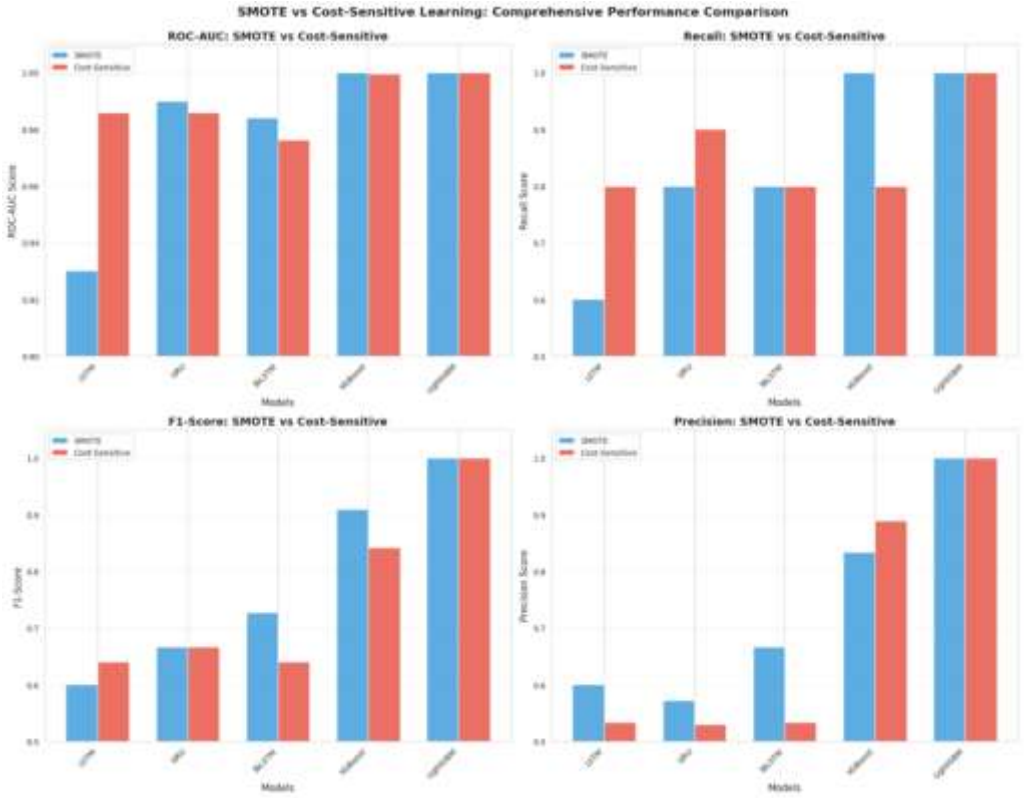

Figure 11. SMOTE vs Cost-Sensitive Comprehensive Performance Comparison

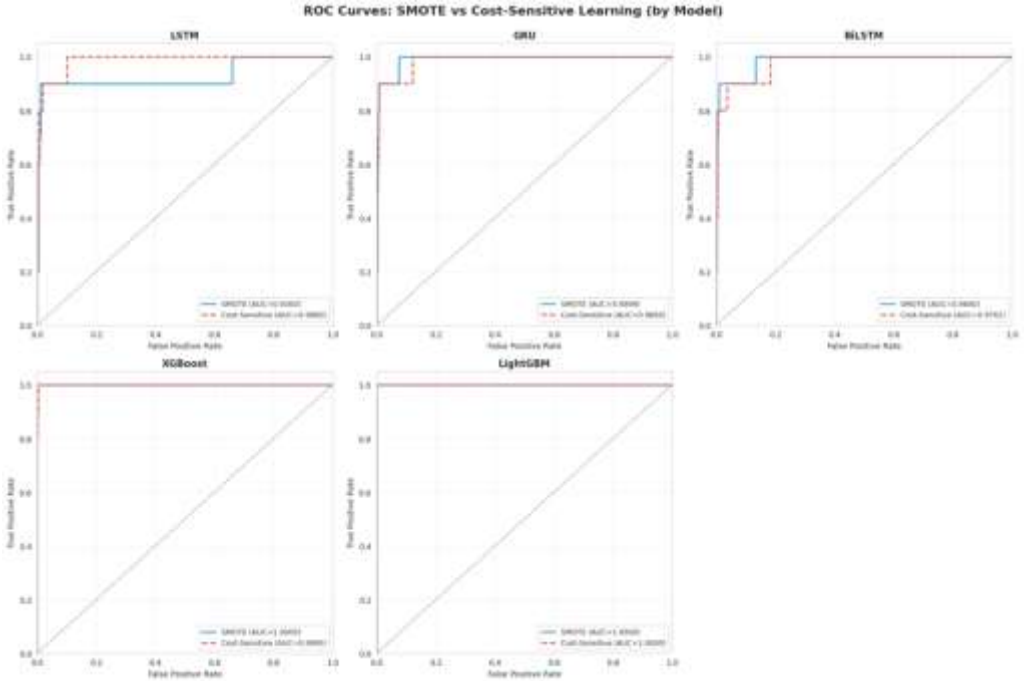

Figure 12. SMOTE - Cost-Sensitive Learning ROC Curves

The modeling phase evaluates RNN-based architectures (LSTM, GRU, and Bidirectional LSTM) alongside tree-based models (XGBoost and LightGBM). Among RNNs, GRU and Bidirectional LSTM outperform unidirectional LSTM, suggesting that simpler gating structures or bidirectional dependency modeling are more suitable for the transaction data. Nevertheless, RNN models generally face challenges in maintaining an optimal balance between detection sensitivity and prediction accuracy.

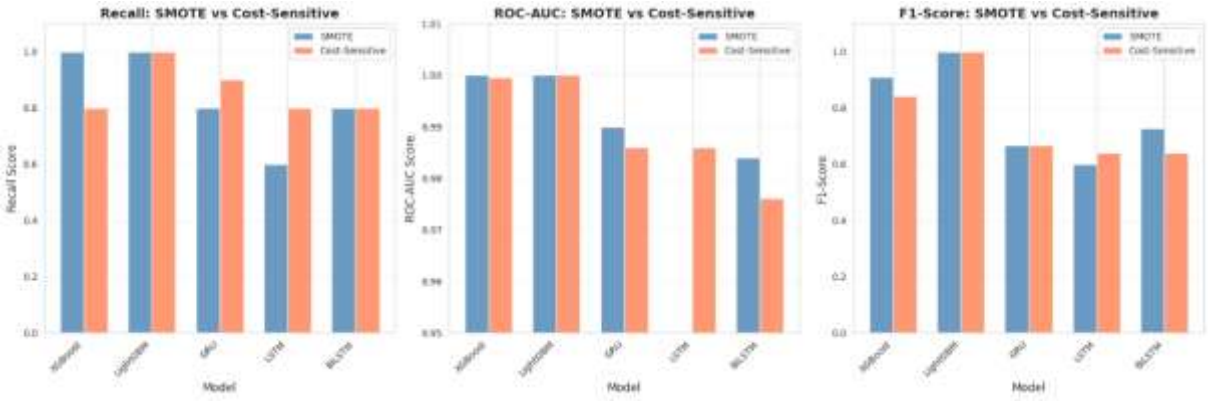

Figure 13. Comparison SMOTE – Cost-sensitive

In contrast, tree-based models demonstrate consistently superior and more stable performance. XGBoost and LightGBM efficiently exploit tabular data characteristics and heterogeneous features, achieving strong results with minimal hyperparameter tuning. Evaluation metrics, including confusion matrices, ROC-AUC, fraud recall, F1-score, and Average Precision, consistently indicate that tree-based models achieve better class separation and more balanced fraud detection outcomes than RNN-based approaches.

Overall, the results indicate that model architecture exerts a greater influence on fraud detection performance than the choice of imbalance handling technique. SMOTE tends to produce better overall balance through higher F1-scores, while cost-sensitive learning is more effective when maximizing fraud recall is the primary objective. Based on comprehensive evaluation, XGBoost combined with SMOTE emerges as the most robust and production-ready solution, offering an optimal trade-off between detection sensitivity, predictive accuracy, and computational efficiency.

Best Model Selection and Final Recommendation

Based on a comprehensive evaluation of all tested models and imbalance handling techniques, XGBoost combined with SMOTE is identified as the most suitable model for fraud detection in this study. This recommendation is grounded in a comparative analysis of primary and supporting evaluation metrics, model robustness, and practical deployment considerations.

From the perspective of the primary metric, fraud recall, XGBoost with SMOTE achieves a perfect score of 1.00, indicating that all fraudulent transactions are correctly detected with no false negatives. This performance is critical in fraud detection, where undetected fraud poses significant financial, regulatory, and reputational risks. The model also attains a ROC-AUC of 1.00, reflecting excellent discriminative capability in separating fraud and non-fraud transactions. In addition, the F1-score of 0.91 confirms a strong balance between precision and recall, ensuring reliable overall classification performance.

When compared with alternative approaches, XGBoost with SMOTE outperforms its cost-sensitive counterpart, particularly in fraud recall, and consistently exceeds the performance of RNN-based models, including the best-performing GRU variant. These results indicate that tree-based models are more effective and efficient for tabular transaction data than sequence-based architectures, which tend to exhibit higher sensitivity to training configuration and greater trade-offs between recall and precision.

From an academic perspective, the recommendation aligns with key principles of fraud detection under imbalanced data conditions. SMOTE effectively enhances minority class representation, allowing the model to better learn fraud patterns, while XGBoost provides strong generalization, stability, and resistance to overfitting. From a practical standpoint, XGBoost offers fast inference, scalability, and model interpretability through feature importance, making it well suited for real-time deployment.

In summary, XGBoost with SMOTE represents the most effective, stable, and production-ready solution in this study. For real-world implementation, continuous performance monitoring, periodic retraining with recent fraud data, and feedback-driven error analysis are recommended. Future work may also explore ensemble strategies to further enhance system robustness and reliability.

5. Conclusion

This study investigates the optimization and comparative performance of Recurrent Neural Network (RNN) models and tree-based models for fraud detection in digital banking under highly imbalanced data conditions. The experimental results demonstrate that class imbalance has a substantial impact on fraud detection performance, particularly by increasing the risk of false negatives when not properly addressed. Therefore, effective imbalance handling is essential to ensure reliable fraud identification. The application of imbalance handling techniques significantly improves model sensitivity to fraudulent transactions. Among the evaluated methods, SMOTE provides more consistent improvements in recall, F1-score, and ROC-AUC, especially when applied to tree-based models. Cost-sensitive learning remains a viable alternative in scenarios where misclassification costs must be explicitly aligned with business risk considerations. Comparative analysis reveals that tree-based models, particularly XGBoost and LightGBM, outperform RNN-based models on tabular transaction data in terms of stability, efficiency, and overall predictive balance. While RNN models are capable of capturing temporal transaction patterns, they require higher computational resources and exhibit greater sensitivity to training configurations. Based on comprehensive evaluation, XGBoost combined with SMOTE and hyperparameter optimization is identified as the most effective model, achieving an optimal trade-off between accuracy, fraud detection sensitivity, and computational efficiency. Overall, this study highlights that the integration of appropriate imbalance handling strategies and model optimization techniques is critical for developing robust and practical fraud detection systems in digital banking environments.

6. References

[1] O. Jasa, “PERATURAN OTORITAS JASA KEUANGAN. REPUBLIK INDONESIA. NOMOR 12 /POJK.03/2021,” 2021, [Online]. Available: https://www.ojk.go.id/id/regulasi/Documents/Pages/Bank-Umum/POJK 12 - 03 -2021.pdf

[2] “2020 ACFE Report to the Nations.” Accessed: Feb. 09, 2026. [Online]. Available: https://legacy.acfe.com/report-to-the-nations/2020/

[3] A. C. Bahnsen, D. Aouada, and B. Ottersten, “Example-dependent cost-sensitive decision trees,” *Expert Syst. Appl.*, vol. 42, no. 19, pp. 6609–6619, 2015.

[4] A. Dal Pozzolo, “Adaptive machine learning for credit card fraud detection,” 2015.

[5] F. Carcillo, A. Dal Pozzolo, Y.-A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, “Scarff: a scalable framework for streaming credit card fraud detection with spark,” *Inf. fusion*, vol. 41, pp. 182–194, 2018.

[6] T. Chen, “XGBoost: A Scalable Tree Boosting System,” *Cornell Univ.*, 2016.

[7] E. Ileberi, Y. Sun, and Z. Wang, “Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost,” *IEEE access*, vol. 9, pp. 165286–165294, 2021.

[8] S. Mittal and S. Tyagi, "Performance evaluation of machine learning algorithms for credit card fraud detection," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, 2019, pp. 320–324.

[9] N. Baisholan, J. E. Dietz, S. Gnatyuk, M. Turdalyuly, E. T. Matson, and K. Baisholanova, "FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets," *Computers*, vol. 14, no. 4, p. 120, 2025.

[10] A. D. Novika and A. Mujhid, "Cost-Sensitive Learning with LightGBM for Class Imbalance in Intrusion Detection Systems," *Eng. Math. Comput. Sci. J.*, vol. 7, no. 2, pp. 147–154, 2025.

[11] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–50, 2016.

[12] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, 2012.

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[14] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[15] A. Mehdary, A. Chehri, A. Jakimi, and R. Saadane, "Hyperparameter optimization with genetic algorithms and XGBoost: a step forward in smart grid fraud detection," *sensors*, vol. 24, no. 4, p. 1230, 2024.

[16] A. Rezaei, M. Yazdinejad, and M. Sookhak, "Credit Card Fraud Detection Using Tree-Based Algorithms For Highly Imbalanced Data," in *2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI)*, IEEE, 2024, pp. 1–6.

[17] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Inf. Sci. (Ny).*, vol. 479, pp. 448–455, 2019.

[18] K. R. L. Reddy, "Advancing Anomaly Detection in Banking Transactions: Leveraging Natural Language Processing and Artificial Neural Network Methods." The George Washington University, 2026.

[19] J. L. S. Saquicela, L. A. B. Herrera, V. P. M. Hidalgo, V. E. C. Cajas, P. D. L. Á. P. Villacis, and V. E. R. Arboleda, "Credit Card Fraud Detection Using Bidirectional LSTM with Attention Mechanism on Sequential Spending Behavior," in *2025 Second International Conference on Intelligent Technologies for Sustainable Electric and Communications Systems (iTech SECOM)*, IEEE, 2025, pp. 1–7.

[20] A. Agarwal, M. Iqbal, B. Mitra, V. Kumar, and N. Lal, "Hybrid CNN-BILSTM-attention based identification and prevention system for banking transactions.," 2021.

[21] P. Feng, "Hybrid BiLSTM-Transformer Model for Identifying Fraudulent Transactions in Financial Systems," *J. Comput. Sci. Softw. Appl.*, vol. 5, no. 3, 2025.

[22] D. Sharma, "A survey of image forensics: Exploring forgery detection in image colorization," 2025.