Fraud Detection in Banking Data By Machine Learning Technique

Mohammed Saniya[1], Md .Arif Mohammad Abdul[2]
[1]Computer Science and Engineering GITAM University Hyderabad, India
[2]Department of Computer Science & Engineering GITAM University, Hyderabad, India

***Abstract*—Real-time fraud detection is crucial for financial institutions because of the sharp increase in fraudulent activity brought about by the quick growth of digital banking and internet transactions. However, there are significant obstacles due to the very unbalanced structure of fraud datasets and the constantly changing fraud patterns. This paper suggests an enhanced fraud detection approach that combines sophisticated ensemble machine learning methods, Bayesian hyperparameter optimization, and class weight-tuning. A majority-voting approach is used to merge LightGBM, XGBoost, and CatBoost models, and deep learning is used to further optimize weight-tuning hyperparameters for unbalanced data.Significant performance gains are seen in experiments using real-world banking data. The suggested approach confirms its resilience despite severe class imbalance by achieving recall improvements of up to 12%, precision improvements of up to 97%, F1-scores above 95%, and MCC values considerably higher than baseline models. When compared to state-of-the-art methods, the optimized framework shows better detection capability, and ensemble combinations like LG+XG+CAT outperform individual models. These findings demonstrate that the suggested hybrid, weight-tuned ensemble technique offers a very successful method for detecting banking fraud in the real world.**

***Index Terms*—Fraud detection using credit cards, Machine Learning Frameworks, Predictive Fraud Systems Deep Learning Enhancement**

I. INTRODUCTION

Global financial transactions have expanded dramatically due to the quick development of digital banking, e-commerce platforms, and online payment systems. Fraudulent activity has increased in tandem with this growth, which presents a significant obstacle for both clients and financial institutions. Particularly, credit card fraud continues to be one of the most common and expensive types of financial crime, causing billions of dollars' worth of damages annually worldwide [1]. Creating effective, scalable, and intelligent detection systems has become essential since scammers constantly change their [2] tactics.

Conventional fraud detection systems mostly rely on rule-based methods, which frequently struggle with large-scale transactional data and are unable to adjust to new fraud trends. This issue is made more difficult by the extremely unbalanced structure of fraud statistics, where fraudulent transactions account for fewer than 1% of all transactions [2]. In these situations, machine learning methods have become effective instruments that can recognize intricate patterns and increase the precision of fraud detection [3].

A range of machine learning and ensemble techniques have been investigated in recent studies to improve fraud detection performance. Promising outcomes have been demonstrated by strategies including AdaBoost, majority voting, Random Forest, Support Vector Machines, and Artificial Neural Networks [[4][5] . Furthermore, time-based aggregation and statistical analysis are two feature engineering techniques that have been successful in enhancing model performance on actual datasets [6].

But there are still a number of difficulties. Numerous current studies rely on oversampling methods like SMOTE, which can raise computing costs and introduce noise. Furthermore, for large datasets, hyperparameter tuning techniques like GridSearchCV and RandomizedSearchCV take a long time [7]. Due to these constraints, more clever and efficient methods are required.

This research presents an advanced fraud detection approach that combines ensemble machine learning models like LightGBM, XGBoost, and CatBoost with class weight-tuning and Bayesian hyperparameter optimization to address these issues. Hyperparameters are further adjusted using deep learning approaches, especially when dealing with skewed data. The suggested method improves overall fraud detection efficiency by lowering false alarms and increasing prediction robustness through the use of majority-voting ensemble techniques.

This work makes the following contributions:

a. A weight-tuning method for managing transaction datasets with significant imbalances.

b. Bayesian optimization provides quicker and more effective tuning of hyperparameters.

c. Light GBM, XGBoost, and CatBoost are combined in an ensemble architecture for better performance.

d. Thorough analysis employing metrics including precision, recall, F1-score, ROC-AUC, and MCC that are best suited for unbalanced data.

The suggested approach considerably outperforms current state-of-the-art models, offering a dependable and scalable solution for credit card fraud detection, according to experimental results on real-world financial datasets.

II. RELATED WORK

Due to the rapid expansion of online transactions and rising financial concerns, credit card fraud detection has been the subject of much research in recent years. The problems of imbalanced datasets, changing fraud patterns, and the requirement for real-time monitoring have been addressed by a number of machine learning and data-driven strategies.

Early research employed statistical and rule-based techniques, but these systems lacked adaptability to new fraud behaviors. Ensemble learning techniques gained popularity as a way to enhance fraud detection performance. Combining several classifiers with AdaBoost and majority voting greatly increases the detection accuracy on both public and real-world datasets, as Randhawa et al. [6] showed. In a similar vein, Feng [3] highlighted how well boosting-based ensemble approaches handle extremely unbalanced financial data.

A key component in enhancing fraud detection has also been investigated: feature engineering. Using the von Mises distribution, [8] developed novel time-based periodic features and demonstrated how adding domain-specific periodic behavior improves fraud identification performance. Their research made clear how crucial it is to create more significant features rather than depending only on sample strategies.

Numerous studies have been conducted on the issue of managing unbalanced datasets. Puh and Brkić [5] observed that appropriate data balancing is crucial for accurate fraud detection when comparing machine learning models like Random Forest, SVM, and Logistic Regression with under-sampling strategies. [9] obtained similar results, analyzing logistic regression, Naïve Bayes, and KNN under random undersampling and concluding that logistic regression works best in skewed data sets.

Additionally, a number of deep learning-based and hybrid approaches have been put forth. A hybrid machine learning architecture that combines many base learners to improve prediction performance was introduced [10]. A hybrid deep learning model for online fraud detection was presented [11], proving that deep neural networks are capable of capturing more intricate characteristics than conventional machine learning models. Using hyperparameter tuning techniques, Taha and Malebary [12] further improved LightGBM and demonstrated its excellent accuracy and computing efficiency.

Additionally, domain-based modeling and sequence mining techniques have been taken into consideration. A sequence mining-based architecture for identifying fraudulent activity in healthcare transactions was created by Matloob et al. [2], demonstrating that mining service sequences can uncover hidden fraud trends. Their approach is applicable to comprehending transactional behavior for credit card fraud, although being in a different field.

Extensive evaluations have also emphasized the shortcomings and advantages of current methods. Kumaraswamy et al. [7] gave a thorough analysis of healthcare fraud detection techniques and talked about the difficulties in using machine learning in actual fraud situations, with a focus on data imbalance and privacy restrictions. In their review of machine learning-based fraud detection, [13] proposed that multilayer models and real-time APIs could be useful in addressing changing fraud strategies.

In general, prior studies indicate that:

1. Single classifiers are outperformed by ensemble models.

2. Feature engineering improves performance dramatically.
3. For accurate fraud detection, unbalanced data processing is essential.
4. Hybrid models and deep learning present encouraging advancements.
5. For real-world huge datasets, effective hyperparameter optimization (beyond GridSearchCV) is required.

III. PROPOSED METHOD

In order to increase accuracy and recall on highly imbalanced financial transaction data, the suggested solution offers an optimal fraud detection framework. The method combines three main elements: a hybrid ensemble of boosting methods, Bayesian hyperparameter optimization, and class weight adjustment.

3.1 Data Processing

Preprocessing of the dataset includes cleaning up missing values, encoding category variables, and eliminating unnecessary fields. There is no oversampling (like SMOTE). Rather, the approach penalizes misclassification of fraudulent transactions more severely by addressing fraud imbalance through class weight assignment.

3.2 Class Weight Adjustment

Class weights are used to boost the influence of minority-class data during training because fraud occurrences are incredibly uncommon. This keeps the original transaction distribution and prevents the creation of synthetic data.

3.3 Bayesian Hyperparameter Optimization

Critical parameters of LightGBM, XGBoost, and CatBoost are tuned using Bayesian Optimization to improve performance. Compared to conventional GridSearchCV, this approach is quicker and more effective, allowing for greater convergence with fewer trials.

3.4 Ensemble Learning Framework

Three separate boosting models are trained: LightGBM, XGBoost,and CatBoost.Majority voting is used to integrate their forecasts, which lowers model-specific variability and increases overall robustness. In addition to ensuring more stability, the ensemble technique reduces false positives and false negatives.

IV. RESULTS AND DISCUSSION

A dataset of actual banking transactions with a significant class imbalance was used to assess the suggested fraud detection system. The suggested majority-voting ensemble was compared to the performance of the individual models, LightGBM, XGBoost, and CatBoost.

4.1 Performance of Individual Models

Due to the highly unbalanced nature of the data, all three boosting techniques performed well, although recall and MCC varied.

i..Strong precision was attained using LightGBM, although recall was marginally poorer.

ii. Strong precision was attained using LightGBM, although recall was marginally poorer

iii.Out of all the models, CatBoost had the best recall.

Even though each model did very well on its own, none of them was able to consistently reach excellent performance across all evaluation metrics.

4.2 Performance of the Ensemble Model

All individual classifiers were outperformed by the suggested LightGBM + XGBoost + CatBoost ensemble.

Rescission rose to 97%, which decreased false alarms.

Recall increased by over 12%, indicating the successful identification of more fraud incidents.

The F1-score was higher than 95%, indicating a good balance between recall and precision.

Even in cases of high imbalance, MCC greatly improved and confirmed steady forecasts.

Strong discrimination capacity was indicated by the ensemble's higher ROC-AUC values.

The majority vote method lessened the model-specific bias present in particular algorithms and helped stabilize forecasts across various transaction patterns.

4.3 Effects of Bayesian Optimization and Class Weighting

Without utilizing oversampling methods, the class weight-tuning strategy enhanced the identification of minority fraud cases.By finding ideal hyperparameters more quickly than grid-based techniques, Bayesian optimization further improved model performance.

In contrast to conventional methods:

Training time reduced,

Generalization improved,

Overfitting decreased.

offers a very successful approach for detecting fraud in the realworld.The suggested approach ensures fewer false positives for banking systems by maintaining high precision while also improving recall, which is essential for fraud detection.When compared to independent machine learning methods, the hybrid model offers a more dependable and scalable solution overall.

V. CONCLUSION

Class weight-tuning, Bayesian hyperparameter optimization, and a hybrid ensemble of LightGBM, XGBoost, and CatBoost are all combined in this study's optimal fraud detection framework. The suggested approach greatly enhances performance when compared to individual models, as shown by the testing results on actual banking data. The ensemble demonstrated its efficacy in identifying fraudulent transactions under severe class imbalance by achieving greater precision, recall, F1-score, ROC-AUC, and MCC.

While Bayesian Optimization offered quicker and more precise parameter tuning, the class weight-tuning approach effectively resolved the imbalance issue without oversampling. Prediction stability and robustness were significantly improved using the majority-voting ensemble. All things considered, the suggested framework offers a dependable, scalable, and high-performance solution for practical fraud detection applications.

In order to improve generalization, future research may concentrate on incorporating real-time detection methods, investigating deep learning architectures, and using the model on multi-bank or cross-institution datasets.

REFERENCES

[1] J. Nanduri, Y. Liu, K. Yang, and Y. Jia, *Ecommerce Fraud Detection Through Fraud Islands and Multi-layer Machine Learning Model*. Springer International Publishing, 2020. doi: 10.1007/978-3-030-39442-4.

[2] I. Matloob, S. A. Khan, R. Rukaiya, M. A. K. Khattak, and S. Member, “A Sequence Mining-Based Novel Architecture for Detecting Fraudulent Transactions in Healthcare Systems,” *IEEE Access*, vol. 10, pp. 48447–48463, 2022, doi: 10.1109/ACCESS.2022.3170888.

[3] I. Sohony, “Ensemble Learning for Credit Card Fraud Detection”.

[4] M. A. Naveed, S. Abdullah, M. M. Ahmed, and B. E. Student, “FRAUD DETECTION IN BANKING DATA BY MACHINE LEARNING TECHNIQUES,” no. 2, 2024.

[5] G. Tong, “Detecting Frauds and Payment Defaults on Credit Card Data Inherited With Imbalanced Class Distribution and Overlapping Class Problems : A Systematic Review,” *IEEE Access*, vol. 12, no. February, pp. 23636–23652, 2024, doi: 10.1109/ACCESS.2024.3362831.

[6] K. Randhawa, C. H. U. K. Loo, and S. Member, “Credit Card Fraud Detection Using AdaBoost and Majority Voting,” *IEEE Access*, vol. 6, pp. 14277–14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[7] B. N. Kumaraswamy, M. K. Markey, T. Ekin, and J. C. Barner, “Healthcare Fraud Data Mining Methods : A Look Back and Look Ahead,” vol. 19, no. 1.

[8] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, “Credit card fraud detection in the era of disruptive technologies : A systematic review,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 1, pp. 145–174, 2023, doi: 10.1016/j.jksuci.2022.11.008.

[9] F. Itoo and M. Satwinder, “¨ ve Bayes Comparison and analysis of logistic regression , Naı and KNN machine learning algorithms for credit card fraud detection,” *Int. J. Inf. Technol.*, 2020, doi: 10.1007/s41870-020-00430-y.

[10] E. F. Malik, K. W. Khaw, B. Belaton, and W. P. Wong, “Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture,” 2022.

[11] M. S. Kumar, “Fraud Detection In Banking Data By Machine Learning Techniques,” vol. 13, no. 4, pp. 429–438, 2025.

[12] A. A. Taha and S. J. Malebary, “An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine,” pp. 25579–25587, 2020.

[13] N. Vinay, G. Rakhi, S. Mahaboob, and R. S. Rao, “Fraud Detection in Banking Data by Machine Learning Techniques,” 2025.