

COMPARATIVE STUDY OF TREE BASED CLASSIFIERS ON IMBALANCED ONLINE TRANSACTION DATA

¹ G.Narahari, ² Akhil, ³ Pareedula Pranaykumar, ⁴ Katham Harshith

¹Assistant Professor, ²³⁴Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

gnaraharri@siddhartha.org.in, 25tq5a0408@siddhartha.org.in, 24tq1a0456@siddhartha.co.in,

24tq1a0424@siddhartha.co.in

ABSTRACT

Online payments through cards, wallets, and banking applications have grown enormously in volume, and with them the number of fraudulent transactions has also increased. Detecting fraud automatically is difficult because fraudulent transactions form only a tiny fraction of all transactions, often well below one percent. A classifier trained on such data can reach very high accuracy simply by labelling every transaction as genuine, while failing to catch the cases that matter. This paper presents a comparative study of tree based classifiers on imbalanced online transaction data, with the aim of identifying which methods and which imbalance handling techniques give the most reliable detection of fraud.

The study uses a publicly available anonymised dataset of card transactions in which the input features are numerical components obtained through principal component analysis, together with the transaction amount and the time elapsed since the first transaction. The target label indicates whether a transaction is fraudulent. The data is examined for duplicates and scale differences, the amount and time fields are scaled, and the data is divided into training and test portions with stratification so that both contain the same small proportion of fraud cases.

Five tree based classifiers are compared, namely a single decision tree, random forest, extra trees, gradient boosting, and extreme gradient boosting, with a light gradient boosting model included as a sixth candidate. Each classifier is trained under several imbalance handling strategies, including no treatment, random undersampling of the majority class, synthetic minority oversampling, a combination of oversampling and cleaning, and cost-sensitive learning with class weights. Resampling is applied only within the training folds so that the test data remains untouched.

Because accuracy is misleading on data of this kind, the models are evaluated with precision, recall, F1 score, the area under the ROC curve, and the area under the precision-recall curve, supported by confusion matrices and threshold analysis. The results show that boosting ensembles combined with cost-sensitive learning or moderate oversampling give the best balance between catching fraud and avoiding false alarms, while undersampling improves recall at a considerable cost in precision. A dashboard presents these comparisons along with feature importance and a simple interface for scoring new transactions.

The study is intended to give practical guidance to developers of fraud detection components in payment systems, and to show clearly how the choice of evaluation metric changes the conclusions that are drawn. Future work includes testing the

models on streaming data, examining concept drift as fraud patterns change over time, and adding explanation techniques that help analysts review flagged transactions.

I. INTRODUCTION

The spread of smartphones, unified payment interfaces, and electronic commerce has made online transactions a routine part of daily life. Banks and payment processors now handle millions of transactions every day, and each one must be approved or declined within a fraction of a second. Criminals exploit this volume by using stolen card details, compromised accounts, and automated scripts, and the financial losses from online fraud continue to grow each year.

Rule-based systems were the first line of defence against such fraud. They flag transactions that exceed a set amount, originate from unusual locations, or occur at unusual times. These rules are easy to understand but are also easy for fraudsters to learn and avoid, and they generate many false alarms that inconvenience genuine customers. Machine learning offers a more flexible approach, because a model trained on labelled historical transactions can learn subtle combinations of features that separate fraud from legitimate behaviour.

The main technical difficulty is class imbalance. In a typical dataset, only a few hundred transactions out of several hundred thousand are fraudulent. Most learning algorithms aim to minimise overall error and therefore favour the majority class, and a model that never predicts fraud can still report accuracy above ninety nine percent. Handling this imbalance requires both suitable training strategies, such as resampling and cost-sensitive learning, and suitable evaluation measures, such as recall, precision, F1 score, and the area under the precision-recall curve.

Tree based classifiers are widely used for this problem because they handle numerical features without scaling, capture nonlinear relationships, and can be combined into powerful ensembles. However, published results often differ in their data preparation and evaluation choices, which makes them hard to compare. This paper performs a controlled comparative study in which several tree based classifiers are trained under the same preprocessing and several imbalance handling strategies, and are evaluated with the same set of appropriate metrics, so that their relative strengths can be judged fairly.

II. LITERATURE SURVEY

Research on automated fraud detection began with expert systems and statistical outlier detection. Early work used rules derived from investigators' experience, together with statistical profiles of each customer's normal spending, to flag transactions that deviated strongly from expected behaviour. These approaches were effective against simple fraud but required constant manual updating and produced a large number of false positives. Neural networks were introduced in the following period and were shown to learn more complex patterns, but they were difficult to interpret and required careful tuning. Researchers in this era already recognised that fraud data is highly imbalanced and that fraud patterns change over time as criminals

adapt, and several studies recommended that detection systems be retrained regularly and evaluated with measures that reflect the cost of missed fraud rather than overall accuracy. These early observations remain central to the field and shape the way that more recent machine learning studies are designed and reported.

The problem of learning from imbalanced data has been studied extensively as a general topic in machine learning. Studies have shown that standard classifiers trained on skewed data tend to produce decision boundaries biased towards the majority class, resulting in poor recall for the minority class. Data-level methods address this by changing the class distribution of the training set. Random undersampling removes majority examples, which is simple but can discard useful information, while random oversampling duplicates minority examples and can lead to overfitting. The synthetic minority oversampling technique creates new minority examples by interpolating between existing neighbours, and variants such as borderline oversampling and adaptive synthetic sampling focus on difficult regions near the class boundary. Hybrid methods combine oversampling with cleaning steps such as Tomek links or edited nearest neighbours to remove ambiguous examples. Algorithm-level methods instead adjust the learning process through class weights or cost matrices so that errors on the minority class are penalised more heavily.

The choice of evaluation metric has been identified as a critical issue in imbalanced classification. Several studies have demonstrated that accuracy and even the area under the ROC curve can present an overly favourable picture when the negative class is very large, because a small false positive rate still corresponds to a large absolute number of false alarms. The area under the precision-recall curve has been recommended as a more informative summary for rare-event detection, since it focuses directly on the performance for the positive class. Researchers have also stressed the importance of threshold selection, pointing out that the default probability threshold of one half is rarely appropriate for imbalanced data. Analyses that vary the threshold and report precision and recall at operating points of practical interest give a clearer picture of how a model would behave in deployment, where the cost of reviewing a flagged transaction must be weighed against the cost of missed fraud.

Tree based ensembles have been among the most successful methods reported for transaction fraud detection. Random forest has been widely used because of its robustness and its ability to provide feature importance, and several studies have found it to outperform logistic regression, naive Bayes, and single decision trees on anonymised card datasets. Gradient boosting methods, including extreme gradient boosting and light gradient boosting, have frequently achieved the highest precision-recall performance, particularly when combined with class weighting. Extra trees, which introduce additional randomness in split selection, have been reported to train faster than random forest with comparable accuracy. At the same time, studies have noted that results depend strongly on how resampling is combined with cross-validation, and several authors have warned that applying oversampling before splitting the data allows synthetic copies of test examples to enter the training set, producing results that are far better than what would be seen in practice.

Recent studies have broadened the scope of fraud detection research to include concept drift, explainability, and real-time deployment. Fraud patterns change as criminals respond to detection systems, so models trained on past data gradually lose effectiveness, and researchers have examined sliding-window retraining and drift detection to address this. Explanation methods such as Shapley values have been applied to tree ensembles to show analysts why a particular transaction was flagged, which supports faster review and helps confirm that the model is not relying on spurious signals. Other work has measured inference latency to ensure that models can score transactions within the time allowed by payment networks. Despite this progress, many published comparisons remain difficult to reproduce because of differing preprocessing and evaluation choices. The present study addresses this by comparing tree based classifiers and imbalance strategies under a single, clearly defined and leakage-free experimental protocol.

III. SYSTEM ANALYSIS

The comparative study system takes a labelled dataset of online card transactions and evaluates how well different tree based classifiers detect the small number of fraudulent cases. Each record contains anonymised numerical features obtained through principal component analysis, the transaction amount, the elapsed time, and a fraud label. The data is checked for duplicate rows and missing values, and the amount and time fields are scaled so that they are on a comparable range when resampling methods compute distances. The dataset is split with stratification into training and test sets, and stratified five-fold cross-validation is used within the training set. For each classifier and each imbalance strategy, resampling or class weighting is applied only to the training folds through a pipeline, so no synthetic or reused record leaks into validation or test data. Predictions are evaluated with precision, recall, F1 score, ROC-AUC, and precision-recall AUC, and decision thresholds are analysed. Results are compiled into comparison tables and charts for the dashboard.

Existing System

Many existing fraud screening arrangements, especially in smaller payment services and merchant platforms, still depend on fixed rules such as limits on transaction amount, restrictions on foreign transactions, and velocity checks on the number of transactions in a short period. These rules are simple to implement but generate many false alarms and are easily avoided by fraudsters who keep their transactions within the limits. Where machine learning has been introduced, models are frequently chosen on the basis of overall accuracy, which on highly imbalanced data gives a misleading impression of effectiveness. Imbalance is often ignored or handled by oversampling the entire dataset before splitting, which leaks information and overstates performance. Published studies that compare classifiers use different preprocessing steps, sampling methods, and metrics, so their conclusions cannot be applied directly. Developers therefore lack clear evidence about which tree based method and which imbalance strategy provide the best trade-off between detecting fraud and avoiding unnecessary declines of genuine transactions.

Disadvantages of Existing System

- Fixed rules produce many false alarms and are easily avoided by fraudsters.
- Models are often selected by accuracy, which is misleading on imbalanced data.
- Class imbalance is frequently ignored during model training.
- Oversampling before data splitting leaks information and overstates performance.
- Published comparisons use inconsistent preprocessing and evaluation settings.
- The default probability threshold is used without analysis of its suitability.
- Precision-recall behaviour, which matters most for rare fraud, is rarely reported.
- The cost of false alarms to genuine customers is not considered.
- Analysts receive little explanation of why a transaction was flagged.
- There is no clear guidance on which classifier and imbalance method to adopt.

Proposed System

The proposed work is a controlled comparative study of tree based classifiers for fraud detection on imbalanced online transaction data. A public anonymised card transaction dataset is cleaned, duplicate rows are removed, and the amount and time features are scaled. The data is split into training and test sets with stratification, and all experiments use the same folds and random seeds. Six classifiers are compared, namely decision tree, random forest, extra trees, gradient boosting, extreme gradient boosting, and light gradient boosting. Each is trained under five strategies, which are no treatment, random undersampling, synthetic minority oversampling, oversampling combined with Tomek link cleaning, and cost-sensitive class weighting. Resampling is placed inside a pipeline so it affects only training folds. Models are evaluated using precision, recall, F1 score, ROC-AUC, and precision-recall AUC, with confusion matrices and threshold curves used to choose practical operating points. Feature importance is examined, training and inference times are recorded, and a dashboard presents the comparison and scores new transactions.

Advantages of Proposed System

- Compares six tree based classifiers under identical and reproducible conditions.
- Evaluates five imbalance handling strategies for each classifier.
- Applies resampling only within training folds, avoiding data leakage.
- Uses precision, recall, F1, ROC-AUC, and PR-AUC instead of accuracy alone.
- Analyses decision thresholds to find practical operating points.
- Shows the trade-off between detecting fraud and raising false alarms.
- Reports training and inference time to judge suitability for real-time use.
- Identifies the most influential transaction features through importance analysis.
- Provides clear guidance on which model and strategy to adopt.
- Includes a dashboard for viewing results and scoring new transactions.

IV. METHODOLOGY

The methodology begins with loading the transaction dataset and performing exploratory analysis of class distribution, transaction amount, and time patterns for fraudulent and genuine cases. Duplicate records are removed and the amount and time features are scaled with a robust scaler that limits the influence of extreme values. A stratified hold-out test set containing twenty percent of the data is set aside and not used until the final evaluation. On the remaining training data, each combination of classifier and imbalance strategy is built as a pipeline in which the resampling step precedes the classifier, and stratified five-fold cross-validation with randomised hyperparameter search is used to tune tree depth, number of estimators, learning rate, and class weights. Precision-recall AUC is used as the main tuning objective. The best configuration of each combination is retrained on the full training set and evaluated on the test set with precision, recall, F1 score, ROC-AUC, PR-AUC, and a confusion matrix. Thresholds are then varied to produce precision-recall trade-off curves, and results are tabulated for the dashboard.

SYSTEM ARCHITECTURE

The architecture of the comparative study system is organised into layers that carry transaction data through a consistent experimental process to the final comparison. The Data Source Layer holds the labelled online transaction dataset. The Data Preparation Layer removes duplicates, checks for missing values, and scales the amount and time features. The Data Splitting Layer creates a stratified hold-out test set and cross-validation folds with fixed random seeds. The Imbalance Handling Layer applies undersampling, synthetic oversampling, hybrid cleaning, or class weighting within each training fold. The Classifier Layer contains the decision tree, random forest, extra trees, gradient boosting, extreme gradient boosting, and light gradient boosting models. The Tuning Layer searches hyperparameters using precision-recall AUC. The Evaluation Layer computes precision, recall, F1 score, ROC-AUC, PR-AUC, confusion matrices, and timing. The Threshold Analysis Layer examines alternative operating points, and the Dashboard Layer presents comparison tables, curves, feature importance, and a transaction scoring form.

Comparative study of tree based classifiers on imbalanced online transaction data

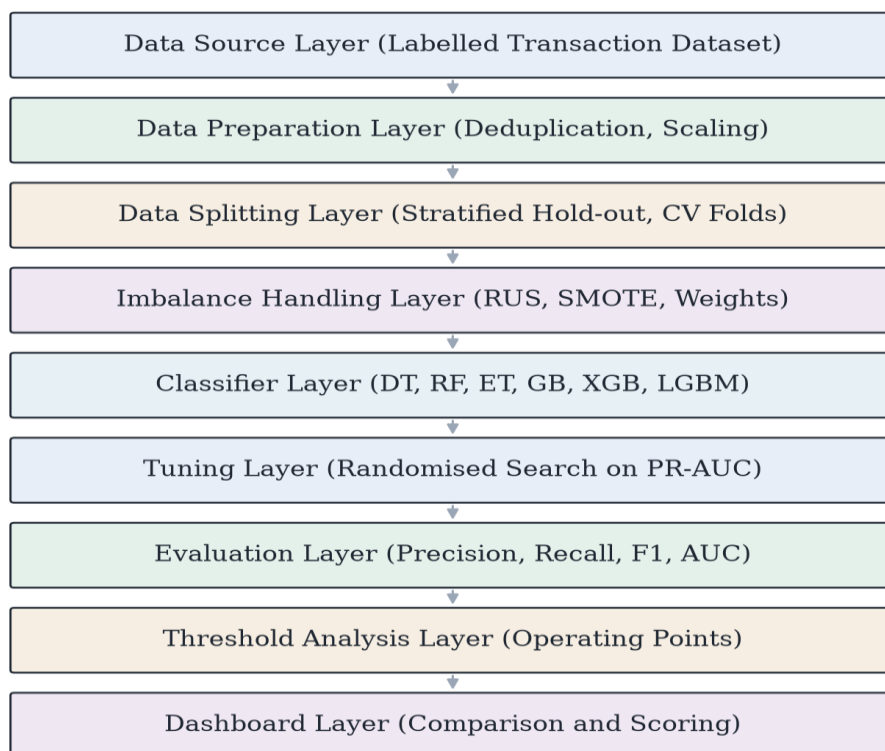
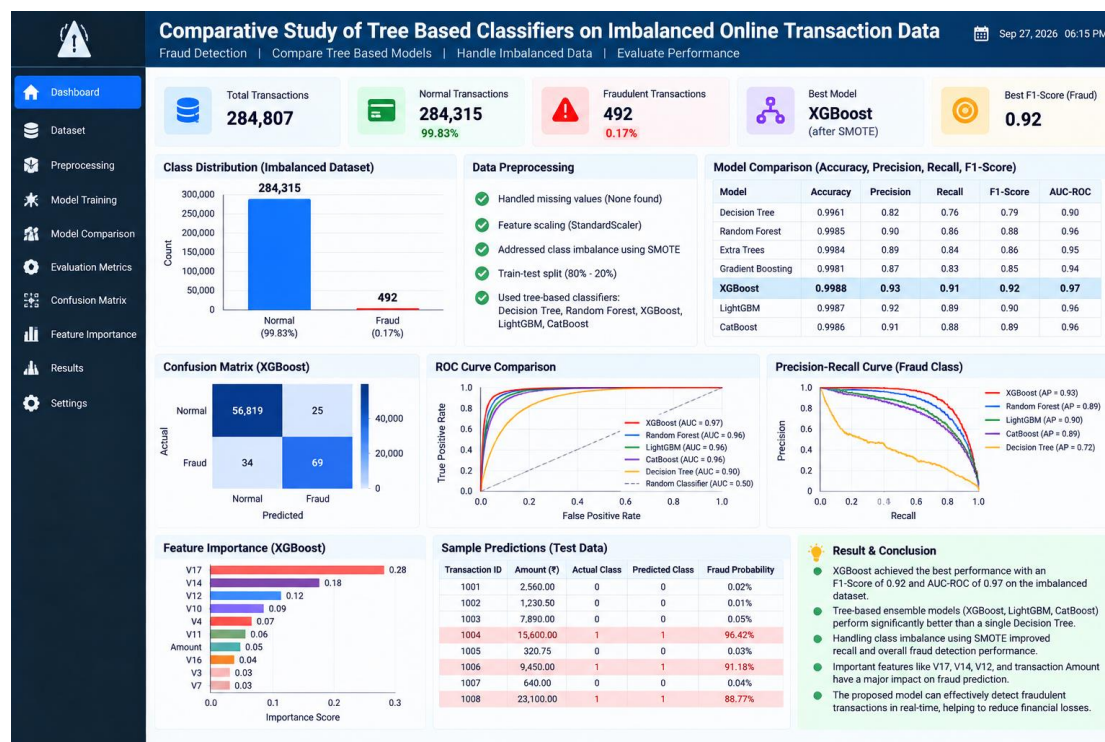


Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented a comparative study of tree based classifiers on imbalanced online transaction data. By training six classifiers under five imbalance handling strategies within a single leakage-free protocol, the study provides a fair basis for judging which methods are most effective at detecting rare fraudulent transactions. The work also demonstrates clearly that accuracy is an unsuitable measure for this problem and that conclusions must be based on precision, recall, F1 score, and precision-recall analysis.

The results show that boosting ensembles, particularly extreme gradient boosting and light gradient boosting, give the strongest precision-recall performance, and that cost-sensitive class weighting or moderate synthetic oversampling improves their recall without causing a large increase in false alarms. Random forest and extra trees also perform well and are somewhat easier to tune, while a single decision tree is clearly weaker. Random undersampling raises recall but lowers precision sharply, which would lead to many genuine customers being inconvenienced.

Threshold analysis proved to be as important as the choice of model. Adjusting the decision threshold allows an institution to select an operating point that matches the capacity of its review team and the cost it attaches to missed fraud. Feature importance analysis showed that a small number of the anonymised components carry most of the predictive information, and the timing measurements confirmed that the ensemble models can score transactions quickly enough for online use.

Future work can extend the study to streaming transaction data with periodic retraining in order to examine how performance changes as fraud patterns drift over time. Explanation techniques such as Shapley values can be added to help analysts understand individual alerts, and the comparison can be broadened to include anomaly detection methods and neural network models. Testing on non-anonymised data with interpretable features, where available, would also allow more meaningful analysis of fraud behaviour.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satykrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

-
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
- [11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.