Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving Fraud Detection Accuracy Through Data Preprocessing and Decision Tree Machine Learning

Darshan Aher

aherdarshan425@gmail.com

Sanjivani University https://orcid.org/0009-0000-1595-9853

Research Article

Keywords: fraud detection, credit card fraud, machine learning, Decision Tree, data preprocessing, class imbalance, SMOTE, feature selection, financial transaction security

Posted Date: October 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11312403/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Improving Fraud Detection Accuracy Through Data Preprocessing and Decision Tree Machine Learning

Darshan Aher
Sanjivani University, Kopargaon, Ahilyanagar, Maharashtra, India

***Abstract*—Digital payments have grown rapidly, and with them the scale and cost of financial fraud. Detection systems in use today still raise many false alarms, struggle with newly emerging fraud patterns, and are hampered by noisy, severely imbalanced transaction data. This paper investigates how far careful data preprocessing can improve the performance of a single, interpretable Decision Tree classifier. Using the public European credit-card dataset (284,807 transactions, of which fewer than 0.2% are fraudulent), the study applies duplicate removal, scaling, a stratified 70/30 split, feature selection, and SMOTE oversampling restricted to the training data. The trained tree (maximum depth 10) was evaluated on an untouched test set that keeps the original class distribution. It reached an accuracy of 0.9897, precision of 0.1152, recall of 0.7746, F1-score of 0.2005, and ROC-AUC of 0.8291, detecting 110 of 142 fraudulent transactions. The results show that a preprocessed Decision Tree can recover most fraud cases at low computational cost and with fully inspectable decision rules, although its low precision means the operating threshold must be tuned to the cost of false alarms.**

***Keywords*—fraud detection, credit card fraud, machine learning, Decision Tree, data preprocessing, class imbalance, SMOTE, feature selection, financial transaction security**

I. INTRODUCTION

The steady shift towards digital payments has been accompanied by a matching rise in financial fraud. Banks and payment processors already operate automated screening systems, yet two weaknesses persist: legitimate transactions are frequently flagged as suspicious, and previously unseen fraud patterns often pass unnoticed. A large share of the difficulty lies in the data. Transaction records can be noisy or incomplete and, above all, are extremely imbalanced, with genuine purchases outnumbering fraudulent ones by several hundred to one. A classifier trained on such data without preparation tends to label nearly everything as genuine and still reports a deceptively high accuracy.

This study asks whether the problem can be tackled mainly through data preparation instead of model complexity. Rather than adopting an ensemble or a deep neural network, it examines how much a disciplined preprocessing pipeline can improve a simple and interpretable Decision Tree. The objectives are (i) to measure the effect of preprocessing on detection performance, (ii) to assess the Decision Tree once the data have been prepared, (iii) to examine false positives and false negatives against a naive baseline, and (iv) to evaluate the model with standard classification metrics.

The remainder of the paper is organised as follows. Section II reviews related work, Section III describes the dataset and methodology, Section IV presents and discusses the results, and Section V concludes with limitations and directions for future work.

II. RELATED WORK

Table I summarises ten recent studies on credit-card fraud detection, listed in approximate chronological order together with the research gap each leaves open.

TABLE I
SUMMARY OF REVIEWED LITERATURE

#	Study	Method	Gap identified
1	Sulaiman et al. [1] (2022)	Survey of DT, RF, SVM, ANN	Little attention to explainable AI and real-time detection
2	Cherif et al. [2] (2023)	Systematic review	Big data, cloud and deep learning need further study
3	Arif et al. [3] (2022)	Hybrid ML model	High computational complexity for real-time use
4	Hondora et al. [4] (2024)	Decision Tree	No comparison of preprocessing methods on imbalanced data
5	Gastkowski et al. [5] (2024)	DT, RF, XGBoost with SMOTE/ADASYN	A single Decision Tree is not optimised
6	Multiple authors [6] (2024)	DT, RF, ANN with SMOTE	Limited evaluation on real-time transaction data
7	AlBulushi et al. [7] (2026)	DT, LR, RF, ANN, XGBoost with SMOTE/ROS/RUS	Lightweight, explainable models need more study
8	Çavdar & Bozanta [8] (2026)	Explainable AI with ML	Limited evaluation across financial settings
9	Gamal et al. [9] (2026)	Hybrid ML and deep learning	High computational cost for real-time use
10	Fazel et al. [10] (2026)	Explainable Boosting Machine	Needs validation on larger real-world banking data

Taken together, the studies show a consistent trajectory. The earlier surveys (2022–2023) mainly compare conventional algorithms such as Decision Tree, Random Forest, SVM and ANN, and repeatedly identify preprocessing and class-imbalance handling as the most pressing issues. Work from around 2024 adopts balancing techniques such as SMOTE and ADASYN and hybrid models; it reports genuine accuracy gains, but at the price of extra computation and without separating the contribution of preprocessing from that of the model. The most recent studies (2026) move towards ensembles,

explainable AI and hybrid deep learning, with growing emphasis on interpretability and deployment.

A common gap emerges: ensemble and deep-learning models usually score slightly higher but sacrifice the transparency and efficiency that a single Decision Tree can offer. This motivates the present work, which tests how far preprocessing alone can lift a lightweight and explainable classifier.

III. METHODOLOGY

A. Workflow

Fig. 1 shows the end-to-end workflow. The transaction data are loaded and cleaned, split with stratification into training and test subsets, balanced on the training portion only, reduced to an informative feature set, and used to train a Decision Tree. The model then classifies the untouched test set, and its errors are analysed with standard metrics and a confusion matrix.

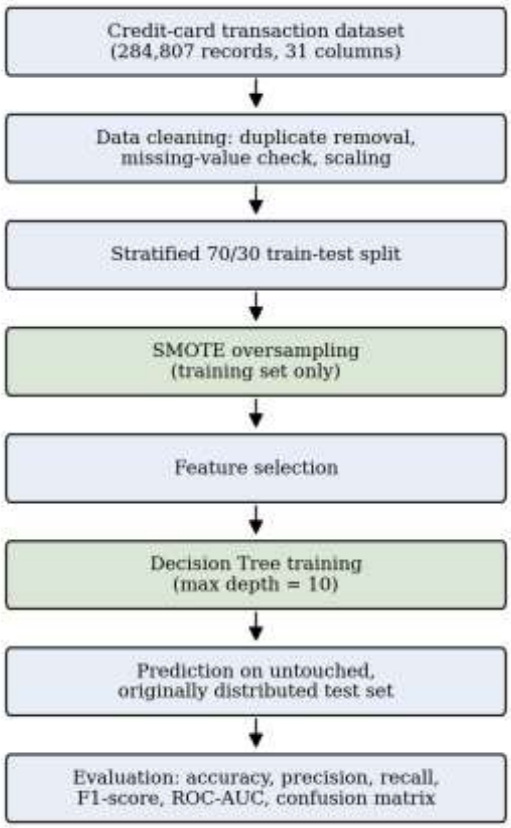


Fig. 1. Research workflow and data pipeline.

B. Dataset

The study uses the publicly available credit-card fraud dataset released by the Machine Learning Group of Université Libre de Bruxelles and Worldline [11]. It records transactions made by European cardholders over two days in September 2013 and contains 284,807 transactions described by 31 columns: the elapsed Time, the transaction Amount, 28 anonymised numerical features (V1–V28) obtained by Principal Component Analysis to protect cardholder identity, and the binary Class label (1 = fraud, 0 = genuine). Only 492 transactions (about 0.17%) are fraudulent, which makes the dataset a standard benchmark for imbalanced classification.

Amount is measured on a ratio scale and Time on an interval scale, while V1–V28 are treated as continuous transformed measurements. A transaction is labelled fraudulent only when it was confirmed as unauthorised or disputed by the cardholder or issuing bank within the dataset's own labelling process.

C. Preprocessing and Sampling

The data were first inspected for missing values and duplicate rows. Removing 1,081 duplicate records left 283,726 unique transactions, of which 473 are fraudulent. Numerical features were then scaled. The cleaned data were divided into 70% training and 30% test subsets using stratified sampling, so that both subsets keep the original fraud-to-genuine ratio; the ratio was verified in each subset.

To counter the class imbalance, SMOTE (Synthetic Minority Oversampling Technique) [12] was applied to the training subset only. SMOTE creates new minority-class examples by interpolating between existing fraud cases, which avoids simply duplicating them. The test subset was never resampled, so that evaluation reflects the true, highly imbalanced distribution encountered in practice.

D. Feature Selection

Not all features contribute equally to fraud detection. Feature selection combined the importance scores produced by the Decision Tree itself (based on the reduction in Gini impurity at each split) with correlation-based filtering to remove redundant attributes. The retained feature set was used for all subsequent experiments.

E. Model and Evaluation Metrics

A Decision Tree classifier (scikit-learn, maximum depth 10, random state 42) was trained on the resampled training data. A decision tree was chosen because it is fast, inexpensive to deploy, and produces human-readable rules that can be audited.

Because accuracy is misleading on imbalanced data (a classifier that labels every transaction as genuine would score above 99%), the model is assessed with accuracy, precision, recall, F1-score and ROC-AUC, together with the confusion matrix. With TP, FP, TN and FN denoting true positives, false positives, true negatives and false negatives, precision = TP/(TP+FP), recall = TP/(TP+FN), and $F1 = 2 \cdot \text{precision} \cdot \text{recall} / (\text{precision} + \text{recall})$. Recall is the primary measure of interest, since it reflects the proportion of real fraud that is caught.

F. Ethical Considerations

The dataset is already anonymised through PCA transformation and contains no personally identifying information. Any real deployment of such a model would still need to comply with data-protection standards such as GDPR and PCI-DSS, including secure storage, restricted access and defined retention periods. Possible sources of bias include class imbalance, sampling bias towards particular regions or periods, delayed or missing fraud labels, and temporal drift as fraud tactics evolve. These were mitigated here through stratified sampling, SMOTE on the training data only, and the use of an interpretable model; in production, periodic retraining on recent data would also be required.

IV. RESULTS AND DISCUSSION

A. Train–Test Split

Table II shows that stratified sampling preserved the class ratio in both subsets: fraud accounts for roughly 0.17% of the training records and of the test records. Training and test distributions are therefore comparable, and the evaluation is not distorted by a mismatch between them.

TABLE II
CLASS DISTRIBUTION AFTER STRATIFIED SPLIT

Subset	Records	Genuine	Fraud
Training (70%)	198,608	198,277	331
Test (30%)	85,118	84,976	142
Total	283,726	283,253	473

B. Model Performance

The Decision Tree was evaluated on the untouched, originally distributed test set. The results are reported in Table III, alongside a naive baseline that labels every transaction as genuine.

TABLE III
PERFORMANCE ON THE TEST SET

Metric	Decision Tree	Naive baseline
Accuracy	0.9897	0.9983
Precision	0.1152	undefined
Recall	0.7746	0.0000
F1-score	0.2005	0.0000
ROC-AUC	0.8291	0.5000

The baseline illustrates why accuracy alone is unsuitable here: it attains 0.9983 accuracy while detecting no fraud at all, a higher figure than that of the trained model. The Decision Tree's accuracy is slightly lower because it deliberately flags some genuine transactions in order to capture fraud. Its recall of 0.7746 means that about 77% of fraudulent transactions were identified. The low precision of 0.1152 shows that roughly one in nine flagged transactions is actually fraudulent, which is the familiar trade-off in imbalanced classification: catching more fraud comes at the price of more false alarms.

C. Confusion Matrix

Fig. 2 gives the confusion matrix for the 85,118 test transactions. The model correctly classified 84,131 genuine and 110 fraudulent transactions, misclassified 845 genuine transactions as fraud (false positives), and missed 32 fraudulent transactions (false negatives). This corresponds to a specificity of 0.9901, that is, a false-positive rate of about 1% of genuine transactions.

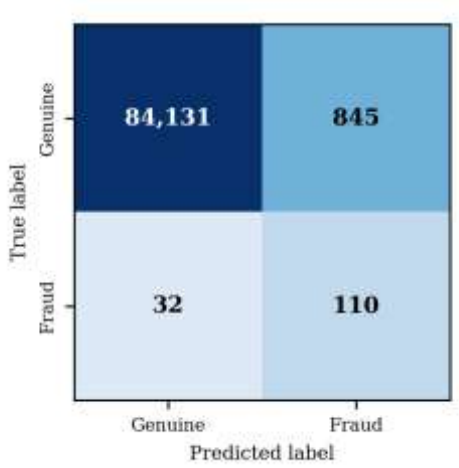


Fig. 2. Confusion matrix on the test set.

D. Discussion

The results answer the research questions in a qualified way. Preprocessing, and in particular SMOTE applied only to the training data, allowed the tree to learn the minority class, so that it detects most fraud instead of exploiting the imbalance. A ROC-AUC of 0.8291 indicates useful but moderate discriminative ability, and the very low precision shows that the model is not yet suitable for fully automated blocking of transactions. In practice it would serve better as a first-stage screening filter whose alerts are reviewed by a second stage or by analysts. The same model is simple, trains quickly and its rules can be inspected, which supports the premise that a carefully prepared dataset lets a single Decision Tree achieve useful detection while remaining lightweight and explainable.

The study has several limitations. The dataset covers only two days of European card transactions, and its features are anonymised, which limits interpretability of individual splits. The comparison with ensemble and deep-learning models is drawn from the literature rather than from experiments under identical conditions, and only one train–test split was evaluated. Finally, oversampling is known to affect precision, and the decision threshold was not tuned.

V. CONCLUSION

This paper examined how much fraud-detection performance depends on data preparation rather than model sophistication. Using duplicate removal, scaling, stratified 70/30 splitting, feature selection and training-only SMOTE, a single Decision Tree reached an accuracy of 0.9897 and a recall of 0.7746 on an untouched, realistically imbalanced test set, correctly identifying 110 of 142 fraudulent transactions. The accompanying precision of 0.1152 reflects the usual precision–recall trade-off and must be balanced against the operational costs of false alarms and missed fraud.

Future work will compare the Decision Tree with Random Forest and XGBoost under the same preprocessing pipeline, quantify the contribution of each preprocessing step through ablation experiments, apply cross-validation and decision-threshold tuning, add explainability techniques so that individual predictions can be audited, and validate the approach on more recent transaction data to guard against concept drift.

REFERENCES

[1] R. B. Sulaiman, V. Schetinin, and P. Sant, “Review of machine learning approach on credit card fraud detection,” 2022.

[2] A. Cherif et al., “Credit card fraud detection in the era of disruptive technologies: A systematic review,” 2023.

[3] M. Arif et al., “Credit card fraud detection using a new hybrid machine learning architecture,” 2022.

[4] E. Hondora, L. Donath, and M. Neamtu, “Automatic card fraud detection based on decision tree algorithm,” 2024.

[5] M. Gastkowski et al., “Comparative analysis of machine learning algorithms and data balancing techniques for credit card fraud detection,” 2024.

[6] Multiple Authors, “Credit card fraud detection using machine learning techniques,” 2024.

[7] A. AlBulushi et al., “Applying supervised machine learning algorithms and ensemble models to enhance credit card fraud detection,” 2026.

[8] E. Z. Çavdar and A. Bozanta, “AI-based credit card fraud detection: A machine learning approach with model explainability on real-world data,” 2026.

[9] N. Gamal, E. M. G. Younis, and W. M. Makram, “Enhancing credit card fraud detection with a hybrid approach using machine and deep learning,” 2026.

[10] R. E. Fazel, A. Bakhtiary, and S. A. Bigdeli, “Improving credit card fraud detection with an optimized explainable boosting machine,” 2026.

[11] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” in Proc. IEEE Symp. Series on Computational Intelligence (SSCI), 2015, pp. 159–166.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.