Journal of Basics and Applied Sciences Research (JOBASR)
ISSN (print): 3026-9091, ISSN (online): 1597-9962
Volume 3(5) September 2025
DOI: https://dx.doi.org/10.4314/jobasr.v3i5.7

Addressing Class Imbalance in Credit Card Fraud Detection with Ensemble Learning and Domain-Specific Feature Engineering

Olaniran, S. F. [1*] & Lawal, M. A.[2]
[1&2]Department of Statistics, Kwara State University Malete, Nigeria
[*]Corresponding Author Email: saidat.olaniran@kwasu.edu.ng

ABSTRACT
Credit card fraud detection remains a critical challenge due to highly class imbalance, changing attack strategies, and the trade-off between recall and precision. This study evaluates the performance of supervised algorithms and ensemble methods (Random Forest, Gradient Boosting Machines (GBM), and Stacking) on a real-world transaction dataset enhanced with temporal, behavioral, and geographic features. A quantitative experimental design was employed, incorporating domain-specific feature engineering and the Synthetic Minority Oversampling Technique (SMOTE) to address imbalance. Models were assessed using precision, recall, F1-score, balanced accuracy, and ROC-AUC. Results show that ensemble models consistently outperformed single classifiers. GBM achieved the highest recall (89.37%), balanced accuracy (94.47%) and ROC-AUC (99.52%) on the imbalanced dataset with engineered features, making it highly effective for minimizing undetected fraud, while Stacking delivered superior precision (95.58%), accuracy (98.90%) and f1-score (92.07%), highlighting its value in reducing false positives. Feature engineering substantially improved recall and balanced accuracy in imbalanced scenarios, while SMOTE enhanced recall for simpler models but sometimes reduced precision. Overall, GBM with engineered features is best suited for real-time fraud screening where recall is critical, whereas Stacking is more appropriate for balanced contexts requiring equal emphasis on recall and precision. These findings underscore the operational value of combining ensemble learning, targeted feature engineering, and imbalance handling to strengthen fraud detection in highly skewed datasets, offering practical guidance for financial institutions seeking more reliable fraud prevention systems.

Keywords:
Credit card, Class imbalance, Ensemble learning, Fraud detection, Feature engineering, SMOTE.

INTRODUCTION

Financial fraud remains a pervasive and costly threat to individuals, businesses, and financial institutions worldwide. The proliferation of digital payment systems, including credit card transactions, has accelerated the frequency and scale of fraudulent activities, with global losses estimated in the tens of billions of dollars annually (Javelin Strategy & Research, 2022; Association of Certified Fraud Examiners, 2023). Credit card fraud, in particular, has emerged as one of the most common forms of financial crime, driven by increased online commerce, data breaches, and sophisticated fraud schemes (Chandola et al., 2021).

Detecting such fraudulent activities presents a unique set of challenges. Foremost among these is the class imbalance problem, wherein fraudulent cases of transactions represent a very small portion of the total transaction volume (Wang et al.,). This imbalance skews model training toward the majority (legitimate) class, leading to elevated false negative rates (instances where fraudulent activities are missed) (Joshi & Malik, 2025). Traditional methods for detecting fraudulent transactions are struggling to keep pace with the changing techniques employed by fraudsters (Chy, 2024).

This study aims to compare selected supervised learning methods on an imbalanced credit card dataset, evaluating their effectiveness using accuracy, precision, recall, F1-score, Balanced-Accuracy & ROC-AUC,

How to cite this article: Olaniran, S. F. & Lawal, M. A. (2025). Addressing Class Imbalance in Credit Card Fraud Detection with Ensemble Learning and Domain-Specific Feature Engineering. *Journal of Basics and Applied Sciences Research*, *3*(5), 52-60. https://dx.doi.org/10.4314/jobasr.v3i5.7

to identify the optimal statistical learning approach for accurately detecting fraudulent credit card transactions in highly imbalanced datasets. The findings will contribute to the discourse on statistical learning methods in fraud detection and offer actionable insights for practitioners

Supervised learning algorithms have been widely applied in fraud detection due to their ability to learn classification boundaries from labeled historical data. Logistic regression is easy to interpret and computationally efficient, but it has difficulty in modeling complex non-linear patterns. without additional feature engineering (Christodoulou et al., 2019). Decision trees can model non-linear relationships and are easy to visualize, but they are prone to overfitting (particularly on noisy datasets) unless techniques like pruning or ensemble methods are applied (Halabaku & Bytyçi, 2024). Support Vector Machines (SVMs) excel in high-dimensional spaces and can model non-linear boundaries through kernel functions, but they require careful optimization of parameters and can become computationally extensive when applied to large datasets (Rezvani et al., 2024).

Ensemble techniques, which combine multiple models to enhance predictive accuracy, have become increasingly popular in detecting fraudulence transactions. Bagging methods like Random Forest reduce variance and improve robustness to noise (Du eta al., 2025). Boosting methods like Gradient Boosting iteratively focus on misclassified instances, often achieving higher precision and recall (Imani et al., 2025). Stacking ensembles leverage the predictions of diverse base learners to train a meta-model, effectively capturing a wider range of data patterns (Mienye & Sun,, 2022). Recent studies have demonstrated that ensembles outperform single classifiers in detecting subtle and complex fraud behaviors, particularly in imbalanced datasets (Herath, 2025; Khan et al., 2024).

The quality of input features plays a critical role in the success of fraud detection models. Domain-specific feature engineering such as constructing temporal features (e.g., inter-transaction time), behavioral profiles (e.g., spending patterns), and spatial indicators (e.g., distance between customer and merchant) can substantially improve model sensitivity to fraud (Iseal et al., 2024; Barnty, 2025). Moreover, hybrid approaches that integrate engineered features with deep learning models achieve superior detection performance in complex fraud scenarios (Yu & Luo, 2025).

The extreme imbalance between fraudulent and legitimate transactions is a core obstacle in fraud detection. Various strategies have been proposed to address this, including oversampling methods like SMOTE (Synthetic Minority Oversampling Technique) (Gupta et al., 2023), under sampling and cost-sensitive learning (Makki et al., 2019), and anomaly detection algorithms such as Isolation Forest and One-Class SVM (Li et al., 2021). Studies combining ensemble learning with imbalance-handling techniques have reported significant gains in detection rates without inflating false positives (Khalid et al., 2024; ResearchGate Master's study, 2024).

While prior studies have addressed elements of supervised learning, ensemble modeling, feature engineering, or imbalance handling individually, few have integrated these components into a unified and systematically evaluated framework. This study advances the state of knowledge by:

1. Integrating domain-specific features (including behavioral, temporal, aggregated and geographic attributes) derived from real-world credit card transaction data.

2. Applying SMOTE to mitigate class imbalance while preserving the distributional characteristics of legitimate and fraudulent classes.

3. Conducting a comprehensive comparison of baseline supervised learning algorithms and advanced ensemble methods, evaluated using multiple metrics that reflect operational realities.

4. Demonstrating empirically how the combined use of engineered features, imbalance handling, and ensemble learning yields significant improvements in detection accuracy, recall, and robustness over baseline approaches.

By bridging methodological advances in machine learning with practical constraints in financial fraud detection, this work contributes actionable insights for both academic research and operational fraud prevention systems.

MATERIALS AND METHODS

Research Design

This study employs a quantitative research method to assess the effectiveness of supervised and ensemble learning methods in detecting fraudulent credit card transactions. The workflow involved data preprocessing, feature engineering, class imbalance handling, model development, and performance evaluation. The design ensured consistent conditions for model comparison, allowing reliable assessment of the contribution of ensemble methods and domain-specific features to fraud detection performance.

Dataset Description

The data employed in this research was obtained from an online repository *https://www.kaggle.com/datasets/kartik2112/fraud−detection*. It contains 555,719 transaction records with 22 attributes covering demographic, temporal, geographic, and transactional information. The target variable,

is_fraud, is binary, with **1** indicating fraudulent and **0** indicating legitimate transactions.

Given the high-class imbalance (fraud cases = 2,145, legitimate = 553,574), a 5% random sample of legitimate transactions was selected, while all fraudulent transactions were retained to avoid loss of critical minority-class information. This resulted in a working dataset of 29,824 records. No missing values were present, ensuring data completeness for modeling.

Key Features:

- Transaction metadata: Transaction date & time, transaction number, unix_time, amount, merchant, category.
- Customer attributes: Gender, account number, first name, last name, street, job, location (city, zip, state, latitude, longitude), date of birth.
- Merchant attributes: Location (latitude, longitude).
- Geodemographics: City population.

Data Exploration

Data exploration was performed to understand data structure, establish anomalies, and detect patterns relevant to fraud detection:

- Class imbalance: Fraudulent transactions represented only 7.19% of the dataset.
- Numerical features: *Transaction amounts and city population* were right-skewed. Geographic coordinates of cardholders and merchants were often highly correlated, but larger distances occasionally signaled fraud.
- Categorical features: Fraud rates varied across merchants and transaction categories, with higher fraud observed in *shopping_net*, *grocery_pos*, and *misc_net*.
- Geographic and demographic indicators: Certain smaller cities and states exhibited disproportionately high fraud rates relative to transaction volume, indicating potential targeted attacks.

Data Preprocessing

Handling Outliers: High-value outliers in the "amt" feature were retained, as they may be indicative of fraud.

Encoding Categorical Variables:

- One-hot encoding for nominal variables such as merchant category.
- Label encoding for ordinal-like variables (e.g., transaction hours).

Data Transformation: Skewed continuous features were *normally transformed*, while all the numerical variables were scaled by employing *Standard scaling* to safeguard comparability across features.

Train-Test Splitting: The dataset was divided into 75% for training and 25% testing, with stratified sampling applied to maintain the initial class distribution.

Addressing Class Imbalance

In addressing class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set, creating additional minority-class records (fraud cases) by interpolating between existing minority instances, thereby balancing the class (Yin et al., 2025).

Algorithm Steps:

i. For each minority record x_j identify its *k-nearest neighbors* from the minority class.
ii. Randomly choose one neighbor x_{ki}.
iii. Generate a new synthetic record using linear interpolation:

$$x_{New} = x_j + \alpha \times (x_{kj} - x_j)$$

Where; $\alpha \sim U(0,1)$ is a random number drawn from a uniform distribution.

iv. Repeat until the desired class balance is achieved.

v.

Figure 1: Class Distribution's Plot Before Balancing

Figure 2: Class Distribution's Plot After Balancing

Feature Engineering

Domain-specific feature engineering was performed to improve the model's capability to identify subtle fraud patterns:

1. Behavioral features: Transaction frequency within defined time windows, average transaction amount per customer, and time since last transaction.
2. Temporal features: Transaction hour, day of week, and working day/weekend
3. Geographic features: Distance between cardholder and merchant using latitude and longitude, and user in same city with merchant
4. Aggregated features: Rolling average transaction amount, transaction frequency per day, and standard deviation of amounts per user.

Model Development

1. Baseline Models:

i. Logistic Regression (LR)

It models the probability that a transaction is fraudulent, given the transaction's features. Mathematically:

$$P(Y = 1/X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{q} \beta_j X_j)}}$$

Where; y (fraud=1, legitimate =0) is the target variable, β_0 is the intercept, X_j is the *jth* predictor and β_j is the *jth* coefficient for *jth* predictor.

ii. Support Vector Classifier (SVC)

It finds a decision boundary that maximizes the margin between classes (fraudulence and legitimate classes)

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{j=1}^{q} \xi_j$$

Subject to:

$$y_i(w.\Phi(X_j) + b) \geq 1 - \xi_j, \qquad \xi_j \geq 0$$

Where; w is the weight vector, b is the bias, ξ_j slack variables for misclassifications, C is the penalty, and $\Phi(.)$ is the kernel mapping function.

iii. Decision Tree Classifier (DTC)

It splits input variables into regions R_m and predicts output variable (fraud) probability within each region:

$$\hat{y}(X) = \sum_{m=1}^{M} c_m . \mathbb{I}(X \in R_m)$$

Where; M is the number of leaf nodes, c_m is the predicted class (fraud or legitimate) in region R_m, $\mathbb{I}(.)$ is the indicator function, and R_m are the terminal nodes defined by a sequence of splitting conditions.

2. Ensemble Models:

i. Random Forest (RF): Bagging-based ensemble of multiple decision trees. The final prediction is the majority vote:

$$\hat{y}_{RF}(X) = mode\{\hat{y}_b(X), \qquad b = 1, 2, \dots, B\}$$

Where $\hat{y}_b(X)$ is the prediction of b^{th} tree.

ii. Gradient Boosting Machine (GBM): It sequentially build tree models; each corrects the errors of the preceding model. At phase j:

$$F_j(X) = F_{j-1}(X) + v.h_j(X)$$

Where; F_{j-1} is the preceding model, h_j is the weak learner, and v is the learning rate. The weak learner minimizes:

$$h_j = \arg\min_h \sum_{i=1}^{n} L\left(y_i, \qquad F_{j-1}(X_i) + h(X_i)\right)$$

With $L(.)$ is the loss function.

iii. Stacking Ensemble: Meta-learning framework combining predictions from base leaners (LR, SVC, RF, and GBM) using logistic regression as the meta-learner.

Let base learners be $\{m_1, m_2, \dots, m_n\}$. Their predictions form new variables:

$$Z_i = \left(m_1(X_i),\ m_2(X_i), \dots, m_n(X_i)\right)$$

The meta-learner (h) produces the final prediction:

$$\hat{y}_{stacked}(X_i) = h(Z_i)$$

Models were implemented in *Python* using scikit-learn libraries.

Model Validation and Hyperparameter Tuning

A *3-fold cross-validation* strategy was employed on the training set to reduce overfitting risk. Hyperparameters were tuned using *grid search* for each model, optimizing for PR-ROC due to the imbalanced nature of the data.

Evaluation Metrics

Given the imbalanced dataset, model performance was assessed using:

i. Accuracy: It measures the proportion of truly classified cases among all cases.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

ii. Precision: It measures the proportion of classified frauds that were truly fraudulent cases.

$$\frac{TP}{TP + FP}$$

iii. Recall: It measures the proportion of true fraud cases correctly detected.

$$\frac{TP}{TP + FN}$$

iv. F1-score: It balances precision and recall.

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

v. AUC-ROC: It measures how well model can separate between fraudulent and non-fraudulent cases across all possible decision thresholds.

$$AUC\_ROC = \int_0^1 TPR(FPR)\, d(FPR)$$

vi. Balanced Accuracy: Average of sensitivity and specificity, mitigating class imbalance bias.

$$\frac{Recall + Specificity}{2}$$

Where;

$$Specificity = \frac{TN}{TN + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

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

TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

These metrics deliver a multi-faceted view of model performance that goes well beyond simple accuracy (Ferrer, 2022; Swaminathan & Tantri, 2024).

RESULTS AND DISCUSSION

The presentations and observations/interpretations of the results of four experimental scenarios: (i) Imbalanced dataset without feature engineering, (ii) Balanced dataset without feature engineering, (iii) Imbalanced dataset with feature engineering and (iv) Balanced dataset with feature engineering were made below

Performance was evaluated using accuracy, precision, balanced accuracy, recall, F1-score, and ROC-AUC.

Table 1. Model performance on imbalanced dataset without feature engineering.

Model	Accuracy	Precision	Balanced Accuracy	Recall	F1 Score	ROC-AUC
LR	0.9650	0.8481	0.8082	0.6250	0.7197	0.9614
SVM	0.9811	0.8776	0.9235	0.8563	0.8669	0.9810
DT	0.9824	0.8932	0.9251	0.8582	0.8754	0.9370
RF	0.9755	**0.9514**	0.8456	0.6940	0.8026	0.9839
GBM	0.9836	0.9075	**0.9266**	**0.8601**	0.8831	0.9654
Stacked	**0.9848**	0.9453	0.9170	0.8377	**0.8882**	**0.9871**

Observation:

The Stacking ensemble achieved the highest accuracy (98.48%), higher balanced accuracy (91.78%), highest F1 Score (88.85%) and ROC-AUC (98.70%), while GBM has the highest balanced accuracy and Recall (92.66% and 86.01% respectively).

Table 2. Model performance on balanced dataset without feature engineering.

Model	Accuracy	Precision	Balanced Accuracy	Recall	F1 Score	ROC-AUC
LR	0.8876	0.3798	0.8887	**0.8899**	0.5324	0.9649
SVM	0.9686	0.7459	0.9160	0.8545	0.7965	0.9731
DT	0.9776	0.8312	0.9251	0.8638	0.8472	0.9251
RF	0.9781	0.8255	0.9340	0.8825	0.8530	0.9904
GBM	0.9850	0.9030	**0.9394**	0.8862	0.8945	0.9858
Stacked	**0.9858**	**0.9300**	0.9312	0.8675	**0.8977**	**0.9918**

Observation:

Balancing increased recall for LR (62.50% → 88.99%) but reduced precision sharply (84.81% → 37.98%). GBM achieved the highest balanced accuracy (93.94%), while Stacked achieved highest in accuracy (98.58%), precision

(93%), F1 score (89.77%) and ROC-AUC (99.18%) with higher balanced accuracy (93.12%).

Table 3. Model performance on imbalanced dataset with feature engineering.

Model	Accuracy	Precision	Balanced Accuracy	Recall	F1 Score	ROC-AUC
LR	0.9749	0.8674	0.8798	0.7687	0.8150	0.9876
SVM	0.9740	0.8783	0.8664	0.7407	0.8036	0.9726
DT	0.9795	0.8998	0.8986	0.8041	0.8493	0.9747
RF	0.9748	**0.9833**	0.8298	0.6604	0.7902	0.9924
GBM	0.9883	0.9441	**0.9447**	**0.8937**	0.9167	**0.9952**
Stacked	**0.9890**	0.9558	0.9424	0.8881	**0.9207**	0.9947

Observation:
Feature engineering improved balanced accuracy and recall for top ensemble models. GBM achieved the highest recall (89.37%), balanced accuracy (94.47%) and ROC-AUC (99.52%), while Stacked achieved the best F1-score (92.07%) and accuracy (98.90%). However, the highest values of metrics across all models and scenarios were achieved among GBM and Stacking in this scenario.

Table 4. Model performance on balanced dataset with feature engineering.

Model	Accuracy	Precision	Balanced Accuracy	Recall	F1 Score	ROC-AUC
LR	0.9484	0.8756	0.6624	0.3284	0.4776	0.9428
SVM	0.9545	0.9227	0.6993	0.4011	0.5592	0.9147
DT	**0.9730**	0.8764	**0.8598**	**0.7276**	**0.7951**	0.9015
RF	0.9453	**0.9923**	0.6203	0.2407	0.3874	0.9792
GBM	0.9688	0.9662	0.7921	0.5858	0.7294	**0.9927**
Stacked	0.9598	0.9720	0.7262	0.4534	0.6183	0.9831

Observation:
Decision Tree delivered balanced results, with the highest balanced accuracy (85.98%), recall (72.76%), F1-score (79.51%), and accuracy (97.30%). GBM also performed strongly, achieving the top ROC-AUC (99.27%) with high precision. Although Random Forest showed exceptional precision (99.23%), its very low recall limited its practical use, confirming Decision Tree and GBM as the most balanced performers.

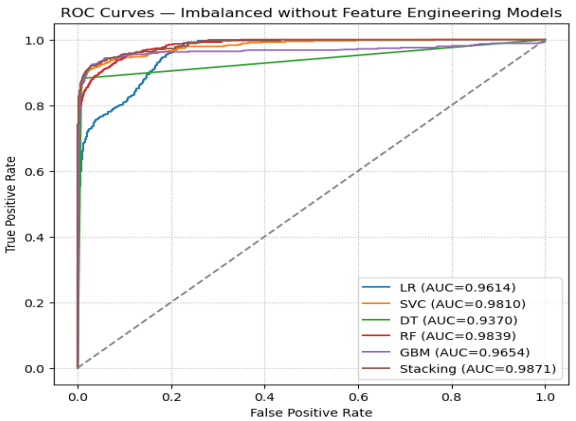

Figure 3: ROC Curves on raw dataset

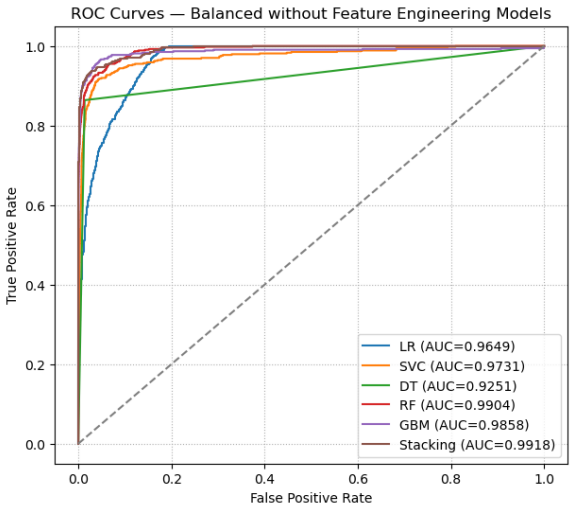

Figure 4: ROC Curves on Balanced dataset without Feature Engineering Models

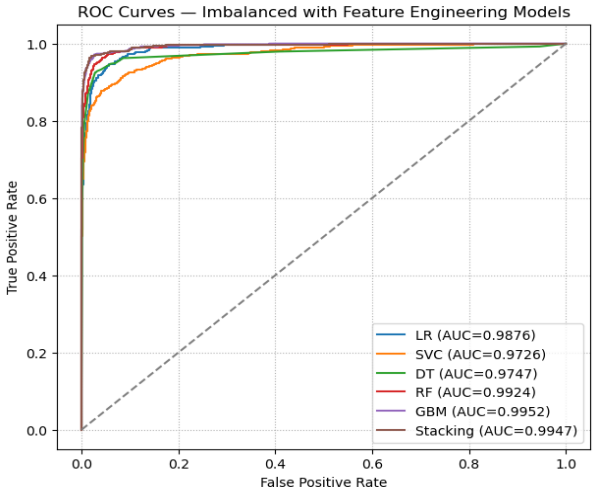

Figure 5: ROC Curves on Imbalanced dataset with Feature Engineering

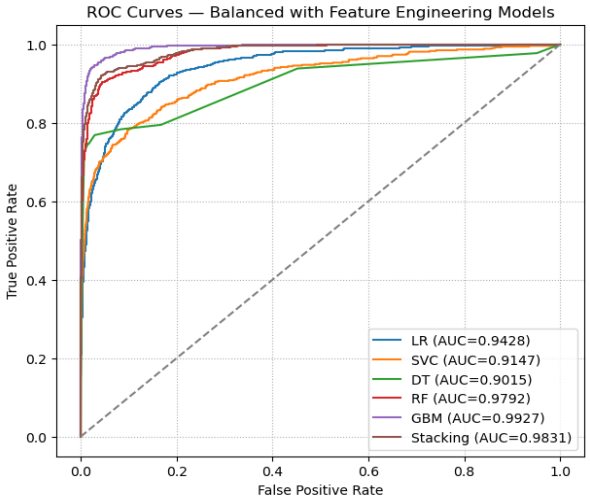

Figure 6: ROC Curves on Balanced dataset with Feature Engineering

Interpretations:

Figures 3–6 show that, without feature engineering, ensemble models (GBM, Random Forest, and Stacking) outperform simpler classifying models (Logistic Regression, Support Vector Classifier and Decision Tree Classifier). With feature engineering, all models improve significantly, with ensemble models achieving near perfect AUCs (>0.99). However, ensemble models benefit from balancing with feature engineering but reduces the performance of simpler models. Overall, feature engineering drives the largest gains, while ensemble methods remain the most reliable for fraudulent cases detection.

The findings of this study underscore the critical role of both feature engineering and ensemble learning in improving the accuracy and robustness of credit card fraud detection. Across all four experimental scenarios, ensemble methods consistently outperformed baseline models. This is consistent with prior studies showing that ensembles can effectively capture complex, non-linear fraud patterns and mitigate the limitations of single classifiers (Herath, 2025; Khan et al., 2024).

The results also reveal that feature engineering provides the greatest performance gains, improving recall, balanced accuracy, and ROC-AUC across almost all models. In the imbalanced dataset with engineered features, GBM achieved the highest recall (89.37%) and ROC-AUC (0.9952), while Stacking achieved the best precision (95.58%) and accuracy (98.90%). These findings highlight the value of incorporating domain-specific behavioral, temporal, and geographic attributes in capturing the subtle signals of fraudulent activity. Similar findings have been reported in other fields, where simple classifier such as SVM and Decision Tree were recommended for prediction tasks, such as predicting user's satisfaction in e-learning system (Imrana et al., 2025).

The role of data balancing was more nuanced. When applied without feature engineering, balancing improved minority class detection and lifted recall, especially for Logistic Regression. However, when combined with feature engineering, balancing benefited ensemble models but reduced the performance of simpler classifiers such as Logistic Regression and SVC. This suggests that while resampling techniques (SMOTE) are valuable for addressing severe class imbalance, their benefits may diminish when rich engineered features already capture fraud-specific patterns. This mixed effect aligns with earlier studies cautioning that oversampling can introduce noise if not carefully tuned (Makki et al., 2019).

From an operational view, these findings have several implications. First, financial institutions should prioritize feature engineering pipelines that incorporate temporal, behavioral, and geographic patterns into fraud detection models. Second, ensemble learning methods, particularly GBM and Stacking, should be favored in production

systems given their superior precision, recall, and ROC-AUC. Third, balancing techniques should be applied selectively, especially when feature-rich datasets are available, as indiscriminate oversampling may harm performance for some models.

Overall, the study demonstrates that integrating ensemble learning, feature engineering, and targeted imbalance handling yields substantial improvements in fraud detection compared to traditional baselines.

CONCLUSION

This study identified the optimal statistical learning approach for detecting fraudulent credit card transactions in highly imbalanced datasets by comparing supervised algorithms and ensemble methods across four scenarios (imbalanced and balanced datasets, with and without domain-specific feature engineering). Ensemble models, particularly Gradient Boosting Machines (GBM) and Stacked ensembles, consistently outperformed single classifiers, while Decision Tree showed competitive results under balanced training with feature engineering. Feature engineering notably improved recall and balanced accuracy in imbalanced datasets, whereas balancing enhanced recall for simpler models but often reduced precision.

Overall, GBM with engineered features on the imbalanced dataset emerged as the most effective configuration for maximizing recall and balanced accuracy, making it well-suited for real-time fraud detection. Stacked ensembles proved more appropriate for balanced datasets or use cases requiring equal emphasis on precision and recall. These findings provide a validated evaluation framework and actionable guidance for operational fraud detection. Future research should explore hybrid approaches combining deep learning with domain-specific feature engineering, and incremental learning mechanisms to further strengthen fraud detection in dynamic and high-volume environments.

REFERENCE

Association of Certified Fraud Examiners. (2023). *2023 report to the nations: Global study on occupational fraud and abuse*.

Barnty, B. (2025). Feature engineering for transaction anomalies. *ResearchGate*. https://www.researchgate.net/publication/

Chandola, V., Banerjee, A., & Kumar, V. (2021). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Chy, M. K. H. (2024). Proactive fraud defense: Machine learning's evolving role in protecting against online fraud. *World Journal of Advanced Research and Reviews, 23*(3), 1580–1589. https://doi.org/10.30574/wjarr.2024.23.3.2811

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Du, K., Zhang, R., Jiang, B., Zeng, J., & Lu, J. (2025). Foundations and innovations in data fusion and ensemble learning for effective consensus. *Mathematics, 13*(4), 587. https://doi.org/10.3390/math13040587

Ferrer, L. (2022). *Analysis and comparison of classification metrics* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2209.05355

Gupta, P., Varshney, A., Khan, M. R., Ahmed, R., Shuaib, M., & Alam, S. (2023). Unbalanced credit card fraud detection data: A machine learning-oriented comparative study of balancing techniques. *Procedia Computer Science, 218*, 2575–2584. https://doi.org/10.1016/j.procs.2023.04.321

Halabaku, E., & Bytyçi, E. (2024). Overfitting in machine learning: A comparative analysis of decision trees and random forests. *Intelligent Automation & Soft Computing, 39*(6), 987–1006. https://doi.org/10.32604/iasc.2024.059429

Herath, H. M. M. N. (2025). Advancing machine learning for financial fraud detection: A comprehensive review of algorithms, challenges, and future directions. *ASEAN Journal of Economic and Economic Education, 4*(1), 49–68. https://ejournal.bumipublikasinusantara.id/index.php/ajeee

Ibanga, J. J. (2024). Resolving data imbalance in financial fraud detection by combining machine learning models and ensemble learning strategies (Master's project). Bournemouth University.

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. *Technologies, 13*(3), 88. https://doi.org/10.3390/technologies13030088

Imrana, S., Obunadike, G. N., & Abubakar, M. (2025). Machine learning-based framework for predicting user satisfaction in e-learning systems. *Journal of Basics and*

Applied Sciences Research, 3(2), 78–85. https://dx.doi.org/10.4314/jobasr.v3i2.9

Iseal, S., Raymond, J., Joseph, O., & Joseph, S. (2024). Financial fraud detection feature engineering techniques for enhanced performance [Paper]. ResearchGate. https://www.researchgate.net/publication/

Javelin Strategy & Research. (2022). *2022 identity fraud study*. https://www.javelinstrategy.com

Joshi, D., & Malik, N. (2025, April 19). Credit card fraud detection using machine learning: Addressing class imbalance with resampling techniques. *SSRN*. https://doi.org/10.2139/ssrn.5225067

Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing credit card fraud detection: An ensemble machine learning approach. *Big Data and Cognitive Computing, 8*(1), 6. https://doi.org/10.3390/bdcc8010006

Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications, 244*, 122778. https://doi.org/10.1016/j.eswa.2023.122778

Li, C., Ding, N., Zhai, Y., & Dong, H. (2021). Comparative study on credit card fraud detection based on different support vector machines. *Intelligent Data Analysis, 25*(1), 105–119. https://doi.org/10.3233/IDA-205317

Luo, S., Ivison, H., Han, S. C., & Poon, J. (2024). Local interpretations for explainable natural language processing: A survey. *ACM Computing Surveys, 56*(9), Article 192. https://doi.org/10.1145/3649450

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access, 7*, 93010–93022. https://doi.org/10.1109/ACCESS.2019.2927250

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access, 10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Swaminathan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research, 27*(4S), 4023–4031. https://doi.org/10.53555/AJBR.v27i4S.4345

Rezvani, S., Pourpanah, F., Lim, C. P., & Wu, Q. M. J. (2024). Methods for class-imbalanced learning with support vector machines: A review and an empirical evaluation. *Soft Computing, 28*, 11873–11894. https://doi.org/10.1007/s00500-024-09931-5

Wang, C., Nie, C., & Liu, Y. (2025). Evaluating supervised learning models for fraud detection. *arXiv*. https://doi.org/10.48550/arXiv.2505.22521

Yin, Y., Zhang, R., & Li, Q. (2025). Resampling approaches to handle class imbalance: A review from a data perspective. *Journal of Big Data, 12*, Article 45. https://doi.org/10.1186/s40537-025-01119-4

Yu, G., & Luo, Z. (2025). Financial fraud detection using a hybrid deep belief network and quantum optimization approach. *SN Computer Science*. https://doi.org/10.1007/s42452-025-06999-y