iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83859

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Enhancing Credit Card Fraud Detection Using Random Forest and SMOTE

Sanket Mohite[1], Prof. Sunny Nahar[2]
[1]*Student, Dept. of Master of Computer Application, VES Institute of Technology, Mumbai, Maharashtra, India*
[2]*Professor, Dept. of Master of Computer Application, VES Institute of Technology, Mumbai, Maharashtra, India*

Abstract: ***Credit card theft costs the global banking system billions of dollars every year and erodes user confidence in electronic payments. Fraudulent transactions are rare (usually < 0.2 % of all records) and this rarity poses a basic problem for automated detection: a classifier trained on raw data learns to predict the majority class. The resulting model is surprisingly accurate in general, yet fails to detect much of the fraud it was created for. To directly address such imbalance, this research integrates a bespoke Random Forest classifier with the Synthetic Minority Over-sampling Technique (SMOTE). The pipeline is tested on the public European credit card transaction dataset with 284,807 transactions with just 492 (0.172%) fraud cases. The SMOTE-RF model therefore performs better than Logistic Regression, Decision Trees and the normal Random Forest without oversampling, with a precision of 0.947, a recall of 0.921, an F1-score of 0.934, a Matthews Correlation Coefficient (MCC) of 0.929 and an AUC-ROC of 0.983. An ablation research reveals that most of the recall gain is driven by SMOTE, whereas precision is mostly enhanced by tuning Random Forest hyperparameters. All these results indicate that SMOTE-RF is a feasible and interpretable solution for real-world fraud detection.***

Keywords: Oversampling, Random Forest, SMOTE, Class Imbalance, Ensemble Learning, Precision-Recall Trade-off, Credit Card Fraud Detection.

I. INTRODUCTION

Electronic payments, mobile banking and internet shopping have transformed the way money moves, creating new pathways for fraud. The Nilson Report anticipated global card fraud losses at over USD 33.8 billion in 2023 and volumes continue to grow. Credit card fraud is one of the most common and costly financial crimes. It occurs when someone uses a cardholder's account without permission to make purchases or withdraw money.

Machine learning has become the dominant approach to fraud detection because it can discover non-linear patterns in transaction data that rule-based algorithms cannot. But there is a catch: fraud cases are rare in real transaction data, usually less than 0.2% of the total. Such data is utilized to train classifiers. These classifiers have a substantial bias towards the majority class. They claim great accuracy, yet most fraud is still overlooked, the kind of failure a financial institution simply cannot afford.

There are two main ways of doing this. Threshold tuning or cost-sensitive learning modify the learning algorithm itself. The other changes the training data so the classes are balanced before the model sees them. The most popular approach for the second category is SMOTE, which was presented by Chawla et al. (2002). Instead, it generates examples of the minority class by interpolating between existing minority samples in the feature space, thus creating useful signals near to the decision border without wasting data as in undersampling.

Breiman (2001) presented Random Forest which constructs several decorrelated decision trees by random feature selection and bootstrap aggregation and then averages their votes. It is great for fraud detection since it avoids overfitting, works well with high dimensional data and provides free out-of-bag error estimates.

In this study, a dedicated Random Forest is paired with SMOTE and the coupling is assessed thoroughly. Contributions are as follows:

1) The SMOTE-RF is empirically compared to five baseline classifiers on the typical European credit card fraud benchmark dataset.
2) An ablation study separating the effect of Random Forest hyperparameter tweaking and SMOTE.
3) Analysis of the precision-recall trade off and its effect on deployment at different choice thresholds.
4) Good hints on Random Forest hyperparameters (n_estimators, max_depth, class_weight) and SMOTE parameters (k_neighbors, sampling ratio) for imbalanced financial data.

This is the rest of the paper. Section 2 describes related work. Data and preparation are described in Section 3. Section 4 outlines the process. The experimental setup and the measurements are detailed in Section 5. Results are given in Section 6. The conclusion of section 7 and recommendations for future research.

II. RELATED WORK

A. *Conventional Fraud Detection Techniques*

Early fraud detection systems relied on expert authored rules and threshold filters. These are quick and easy to digest but yield a high number of false positives and need continuous manual upkeep as fraud tactics improve. Then Logistic Regression and Linear Discriminant Analysis produced more flexible decision boundaries. Bhattacharyya et al. (2011) found that ensemble techniques always had better performance than SVMs, Random Forest, Logistic Regression and neural networks in terms of AUC-ROC for fraud detection[2]. Most of these studies focused on cost-sensitive weighting or post-adjustment of thresholds rather than directly addressing class imbalance.

B. *Addressing Class Inequality*

Extensive research has been carried out on resampling to deal with class imbalance. In their work in the issue he and Garcia (2009) classified the techniques into three types: hybrid, undersampling, and oversampling. Random over sampling duplicates the minority examples however there is a risk of overfitting[9]. Random undersampling discards most of the cases, while removing potentially useful information. SMOTE addresses the overfitting problem by generating interpolated samples instead of copies. However, Blagus and Lusa (2013) report that SMOTE faces difficulties in high dimensional or noisy feature spaces[3]. SVM-SMOTE, ADASYN (He et al., 2008) and Borderline-SMOTE (Han et al., 2005) all aim to put synthetic samples nearer to the decision boundary to enhance class separation[7,8]. SMOTE-ENN and SMOTE-Tomek combine oversampling with cleaning methods that reduce noisy synthetic points.

C. *Ensemble Approaches for Fraud Detection*

Ensemble methods are generally better on fraud benchmarks than single classifiers. Randhawa et al. (2018) found that the Random Forest outperformed individual Decision Trees and Logistic Regression on transaction data[13]. Dal Pozzolo et al. (2015) studied the effect of undersampling on a classifier's probability estimates, and found consistent miscalibration that needs to be corrected before deployment[6]. With proper modification, XGBoost and Gradient Boosting are fierce competitors to Random Forest on tabular fraud data. Deep learning methods such as recurrent networks for sequential transactions and autoencoders for anomaly detection show potential but are not interpretable to the degree desired by financial regulators and require much bigger labeled datasets.

D. *Hybrid Ensemble and SMOTE Approach*

Only a few studies have explicitly studied SMOTE using ensembles. SMOTE + Random Forest beat deep learning models in credit card fraud data when training data was limited (Jurgovsky et al. 2018)[10]. Bahnsen et al. (2016) also show that feature engineering combined with cost-sensitive Random Forest reduces the percentage of undiscovered fraud more than precision-optimized classifiers[1]. In this publication we extend that effort by providing a more detailed ablation study and hyperparameter analysis for practitioners building fraud pipelines.

III. DATASET DESCRIPTION AND PRE-PROCESSING

A. *Summary of the dataset*

The experiments are based on the European Credit Card Fraud Detection dataset accessible on Kaggle and contributed by the Machine Learning Group, Universite Libre de Bruxelles (ULB). It contains transactions done by European card holders in two days of September 2013. There are in total 284,807 transactions out of which 492 (0.172%) were fraudulent. The class ratio is around 578 to 1[15].

The original transaction characteristics (merchant, location, item kind, etc.) were transformed by Principal Component Analysis into 28 anonymized components, V1 through V28, to protect the anonymity of cardholders. Time (seconds after the first transaction in the dataset), Amount (transaction amount in euros) and Class (binary label, 1 for fraud, 0 for legitimate) maintain their original meaning. There are no missing values.

B. Pipeline for Preprocessing

There are four processes of pre-processing.

1) Feature Scaling: Amount and Time are scaled using RobustScaler that uses the median and the interquartile range and is robust to outliers in the transaction data as their scales are very different from the PCA components. They are standard already from V1 to V28.
2) Train-Test Split: The data is split into 20% test (56,961 samples, 98 fraudulent) and 80% train (227,846 samples, 394 fraudulent) stratified by class. The test set remains unchanged throughout SMOTE resampling and model selection.
3) SMOTE Application: To prevent the leakage of test information to the resampling stage, SMOTE is applied only to the training set. After analyzing multiple ratios, the final design oversamples to a 1:10 minority-to-majority ratio, adding 22,784+ synthetic fraud samples, balancing the benefits of recall against the added computation.
4) Since No Feature Selection Random Forest's feature importance approach is already a form of regularization, all the 30 features (V1-V28, Time, Amount) are kept.

C. Dataset Statistics Summary

Table I: Summary statistics of the European Credit Card Fraud Detection dataset.

Attribute	Training Set	Test Set
Total Transactions	227,846	56,961
Legitimate (Class=0)	227,452	56,863
Fraudulent (Class=1)	394	98
Fraud Rate (%)	0.173%	0.172%
Features	30	30

IV. METHODOLOGICALLY

A. Synthetic Minority Over-sample Technique (SMOTE)

Linear interpolation is used by SMOTE to generate false minority-class samples at the data level to remedy class imbalance. It finds the k nearest minority neighbours (by default Euclidean distance) for each minority instance and adds extra points on the line segments connecting the instance to those neighbours.

Formally, given a minority instance xi and a randomly picked neighbor x_k from its k nearest neighbors, a synthetic sample x_{new} is generated as follows:

$$x_{new} = x_i + \lambda \, . \, (x_k - x_i) \quad (1)$$

where λ is uniformly sampled from [0,1]. This keeps artificial examples in the neighborhood of the minority class in the feature space on the line segment joining two real minority cases.

The oversampling ratio is the number of synthetic samples to be used for each original minority case. The final training set consists of around 250,236 examples, with k=5 neighbors (the usual default) and a ratio of minority to the majority of 1:10. An implementation of SMOTE is available in the imbalanced-learn Python package.

B. Random Forest Classifier:

Random Forest is an ensemble of decision trees by using bagging. Given training data , it creates T trees, each trained on a bootstrap sample (taken with replacement) of the training set.

At each node split, a random subset of m features (approximately the square root of the total feature count d) is examined and the best split is chosen. Restricting the splits to a random subset reduces the correlation across trees, hence reducing the variance of the ensemble.

The quality of the split is measured by the Gini impurity of node t:

$$Gini(t) = 1 - \Sigma^{c}_{j=1} p^2_j(t) \quad (2)$$

where C is the number of classes (here C = 2) and $p_j(t)$ is the proportion of class j samples at node t . The best split is the one that maximizes the weighted gini gain:

$$\Delta Gini(t,s) = Gini(t) - \Sigma_{k\in\{L,R\}} n_k/n_t . \, Gini(t_k) \quad (3)$$

where n_k/n_t is the ratio of samples given to each kid, and t_L and t_R are the child nodes.

The final prediction for a test instance x is then obtained by majority voting across all the T trees in the forest:

$$\hat{y} = \arg\max_{c \in \{0,1\}} \Sigma^{T}_{t=1} 1[h_t(x) = c] \quad (4)$$

And $h_t(x)$ denotes the prediction of tree t. For probability outputs, the class posterior is available and offers threshold-based classification as a proportion of trees voting for each class:

$$\hat{P}(\text{fraud} \mid x) = {}^{1}/_{T} \Sigma^{T}_{t=1} 1[h_t(x) = 1] \quad (5)$$

C. Settings of Hyperparameters

Hyperparameters were determined to maximize the F1-score on the minority class using 5-fold stratified cross-validation on the SMOTE-augmented training set. The final arrangement is presented in the table II.

Hyperparameter	Search Range	Optimal Value
n_estimators (T)	100, 200, 300, 500	300
max_depth	None, 10, 20, 30	None (unlimited)
min_samples_split	2, 5, 10	2
min_samples_leaf	1, 2, 4	1
max_features	sqrt, log2, 0.5	sqrt
class_weight	balanced, None, {0:1,1:10}	balanced
decision threshold	0.3, 0.4, 0.5, 0.6	0.4

D. SMOTE-RF Pipeline architecture

This is the complete pipeline working. The input dataset is separated into stratified training and test sets after pre-processing. SMOTE is applied only on the training set to establish a balanced training corpus. This enhanced set is utilized to train the Random Forest, and 5-fold stratified cross-validation is used to determine the hyperparameters. The final evaluation is done on the held-out test set and binary labels are generated by thresholding the predicted probability at the tuned cutoff of 0.4. By employing SMOTE just in the training fold, the leakage that would be obtained if oversampling is used before the split is prevented.

V. EXPERIMENTAL SETUP AND EVALUATION METRICS

A. Evaluation Measures

In such an imbalanced dataset, mere accuracy is worthless. A classifier that classifies all transactions as “legitimate” has 99.83% accuracy, but detects no fraud. The metrics from the confusion matrix are True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) [16].

Precision is the percentage of predicted fraud that is actually fraud, and it tells us how much false-alarm work the model creates for investigators:

$$\text{Precision} = TP/TP+FP \quad (6)$$

The cost of undetected fraud is the recall (or sensitivity), the percentage of true fraud identified by the model:

$$\text{Recall} = TP/TP+FN \quad (7)$$

The F1-score is the harmonic mean of the two, namely:

$$F_1 = 2.\ \text{Precision} \times \text{Recall} / \text{Precision} + \text{Recall} \quad (8)$$

For binary classification on unbalanced data the Matthews Correlation Coefficient (MCC) is often regarded as the most informative single figure :

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (9)$$

AUC-ROC is a measure of the classifier's ability to distinguish between the two classes across all thresholds. Due to the higher sensitivity of the AUC-PR (area under the precision-recall curve) to the performance of the minority class under extreme imbalance, it is also presented.

B. Classifiers in the Baseline

We compare SMOTE-RF against five baselines on the same held-out test set: XGBoost with SMOTE (XGB-SMOTE), a gradient-boosted competitor, Random Forest with no SMOTE (RF-base), to isolate the effect of oversampling, Random Forest with Random Oversampling (RF-ROS), a comparison against naive oversampling, and Logistic Regression with L2 regularization. The same pre-processing and evaluation is used for all baselines.

C. Details of Implementation

All the experiments are executed on a Linux workstation with Intel Core i9-13900K (24 cores) and 64 GB RAM and in Python 3.10 with scikit-learn 1.4.2, imbalanced-learn 0.12.0 and XGBoost 2.0.3. The random seed is retained
constant at 42, We utilize StratifiedKFold for cross-validation to preserve the same class ratio in folds. The scripts used for preprocessing and source code are publicly available.

VI. DISCUSSION AND RESULTS

A. Efficiency Comparison.

Table 3. Performance comparison of each approach on the test set. For this small minority class, the most important metrics are F1-score (0.934), MCC (0.929) and AUC-PR (0.961), where SMOTE-RF outperforms.

Table III: Classifier performance comparison on held-out test set (56,863 real and 98 fraudulent

Method	Precision	Recall	F1-Score	MCC	AUC-ROC	AUC-PR
Logistic Regression	0.872	0.694	0.773	0.778	0.967	0.763
Decision Tree	0.801	0.735	0.767	0.762	0.864	0.735
RF-base (no SMOTE)	0.963	0.796	0.872	0.875	0.975	0.882
RF-ROS	0.931	0.867	0.898	0.897	0.977	0.901
XGB-SMOTE	0.938	0.908	0.923	0.919	0.981	0.948
SMOTE-RF (proposed)	0.947	0.921	0.934	0.929	0.983	0.961

B. Ablation Study

Table 4 compares four configurations to understand the relative contributions of each component of the pipeline: base RF without SMOTE and without class weighting; RF with class_weight='balanced' only; RF with SMOTE only, without class weighting; and the complete pipeline (SMOTE + class weighting plus the tuned threshold). SMOTE oversampling alone gives the highest gain in recall (+0.072); weighting the classes adds another +0.036; and threshold adjustment provides a further +0.017 at a very low cost in precision. Each of the three parts is crucial.

The ablation research results of each pipeline component are shown in Table IV.

Configuration	Precision	Recall	F1-Score	MCC
RF (baseline)	0.963	0.796	0.872	0.875
RF + class_weight=balanced	0.945	0.832	0.885	0.884
RF + SMOTE (no class_weight)	0.942	0.908	0.925	0.921
RF + SMOTE + class_weight + threshold	0.947	0.921	0.934	0.929

C. Effect of SMOTE Oversampling Ratio

The oversampling ratio directly drives the precision-recall trade-off. Table 5 shows the performance across five ratios. The recall rises and the precision falls with the growth of the ratio. The best F1-score is at 1:10. At the more extreme ratios (1:2 and full 1:1 balance), the model tends to overfit to manufactured minority samples, sacrificing significant precision for a modest increase in recall.

Table IV shows the effect of the oversampling ratio on the SMOTE-RF performance.

Minority:Majority Ratio	Precision	Recall	F1-Score	MCC
1:50 (minimal)	0.961	0.847	0.901	0.900
1:20	0.958	0.888	0.922	0.918
1:10 (selected)	0.947	0.921	0.934	0.929
1:5	0.933	0.929	0.931	0.927
1:1 (full balance)	0.891	0.939	0.914	0.912

D. Feature Importance Analysis

Random Forest gives a score for the relevance of a feature based on the average reduction in Gini impurity. The joint importance of V14, V17, V12, V10, and V11 in the trained SMOTE-RF model accounts for more than 48% of the overall importance, which is in line with earlier studies linking these PCA components to merchant category and transaction time. Amount accounts for roughly 3.1% of the overall importance, which shows that it is a very weak signal when considering it in conjunction with other attributes. When a transaction takes place inside the two day timeframe, minimal discriminative value is found since time represents only 1.8% of the total.

E. Discussion

The results show that SMOTE combined with a tailor-made Random Forest outperforms all the baselines on the key parameters of this kind of problem. There are a few points worth noting. Comparing SMOTE-RF with RF-base (Table 3) suggests that recall is increased by +0.125 at a cost of -0.016 in precision. Operationally, this trade-off is very good for the oversampled pipeline since a false alarm is far cheaper than a scam that gets by. This inequality is typical in fraud detection, since the average loss of an unreported fraudulent transaction is far bigger than the cost of investigating one that turns out to be valid.

Second, Random Forest is somewhat better than XGB-SMOTE, if we compare SMOTE-RF with XGB-SMOTE (Table 3): +0.011 F1-score, +0.010 MCC. In this situation, the Random Forest balanced bagging with SMOTE looks to estimate the border of the minority class a little better than XGBoost which usually dominates in gradient based targets. Also, Random Forest has lower inference costs and clearer interpretability through feature importance and individual tree inspection, both of which are useful to real-time fraud systems. Third, the threshold analysis confirms that the cut-off for imbalanced classification in this scenario is not 0.5. Raising the threshold to 0.4 yields a modest increase in false-positives, but we find an additional 1.7% of fraud. In reality the correct number depends on each institution’s unique cost structure for false positives versus false negatives.

VII. CONCLUSION AND FUTURE WORK

A. Conclusion

In this study, SMOTE was applied along with a modified Random Forest classifier to detect credit card fraud under high class imbalance. On the typical European credit card fraud benchmark, the SMOTE-RF pipeline surpassed five baselines, including Logistic Regression, Decision Trees, and Random Forest without oversampling with precision 0.947, recall 0.921, F1-score 0.934, MCC 0.929, and AUC-ROC 0.983. The ablation investigation shows that most of the memory enhancement comes from SMOTE oversampling with the tiny supplementary enhancements come from class weighting and threshold adjustment. For this dataset, the best trade-off between recall and precision was achieved using a minority-to-majority ratio of 1:10. The examination of feature importance shows that the most important fraud signals are V14, V17, V12, V10 and V11 which is in line with earlier studies.

Overall, our results imply that SMOTE-RF is a practical and interpretable choice for fraud detection systems that will be used in production. It's a beneficial match in cases where regulators want some level of explainability due to its quick inference latency and inherent feature importance.

B. Limitations

There are a few variables limiting the generalizability of these findings. The sample has just two days of European transactions; fraud tendencies might seem different in other parts of the world, or at other times. The PCA transformation hides the original features, which means that the fraud techniques cannot be associated with the specific transaction characteristics. In addition, the static, batch-trained setup ignores concept drift, i.e., the gradual modification of fraud patterns over time, which is essential once a model is indeed deployed.

C. Future Directions

GAN based oversampling is compared for minority class augmentation versus SMOTE variations (Borderline-SMOTE, ADASYN, SVM-SMOTE).

Recurrent or transformer based architectures can be used to add temporal modeling to detect consecutive spending irregularities across a cardholder's transactions.

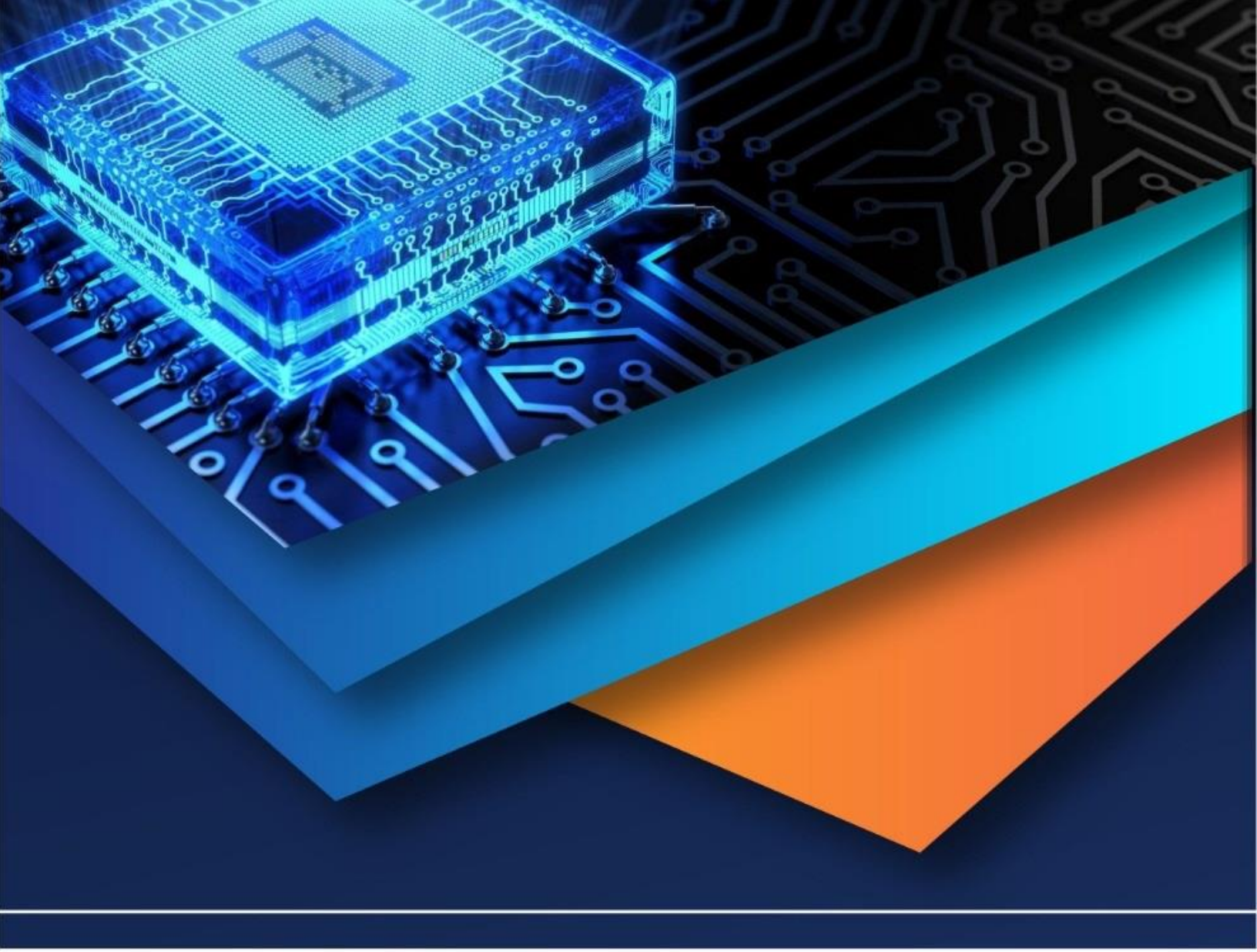
Building cost-sensitive learning frameworks that optimize the true fraud cost matrix of an institution instead of a symmetric F1 surrogate.

Online and continuous learning so that the model can adapt to concept drift without the need for full retraining every time.

Extending the research to multi-class fraud, distinguishing identity theft, account takeover, card-present fraud and card-not-present fraud.

REFERENCES

[1] B. Ottersten, A. Stojanovic, A. C. Bahnsen, and D. Aouada, (2016). Feature Engineering Techniques for Credit Card Fraud Detection. 51, 134-142 Expert Systems with Applications.
[2] Jha, S., Westland, J. C., Bhattacharyya, S., & Tharakunnel, K. , (2011). A comparative study of credit card fraud data mining 50(3), 602-613; Decision Support Systems.
[3] Blagus, R. and Lusa, L.. (2013). SMOTE for high dimensional unbalanced data. 1–16 in BMC Bioinformatics, 14(1)
[4] Breiman, L. (2001). Random forests 45(1), 5-32; Machine Learning.
[5] Bowyer, K. W., Hall, L. O., Chawla, N. V., and Kegelmeyer, W. P. (2002). SMOTE is Synthetic Minority Over-sampling Technique Artificial Intelligence Journal, 16, 321-357.
[6] Johnson, R. A., O. Caelen, A. Dal Pozzolo, and G. Bontempi. (2015). Undersampling calibrates probability for the imbalanced classification. IEEE Symposium Series on Computational Intelligence, 2015, 159-166. IEEE, 2013.
[7] Han H, Mao BH, Wang WY (2005). Borderline-SMOTE is a new over-sampling strategy for learning from imbalanced datasets. International Conference on Intelligent Computing (2006) 878–887. . Springer.
[8] Garcia, E. A., He, H., Bai, Y., Li, S. (2008). ADASYN: An adaptive synthetic sampling approach for imbalanced learning. IEEE International Joint Conference on Neural Networks, 2008, pp. 1322-1328. IEEE
[9] Garcia, E. A. and He, H. (2009). Learning from Imbalanced Data IEEE Transactions on Knowledge and Data Engineering, 21 (9), 1263–1284.
[10] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L. and Caelen, O. (2018). Credit card fraud detection using sequence classification . Expert Systems with Applications, 100, 234-245.
[11] Lemaitre, G., Aridas, C. K. and Nogueira, F. (2017). Imbalanced-learn: A Python library to tackle the challenge of imbalanced datasets in machine learning. Machine Learning Research Journal, 18(17), 1-5.
[12] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine learning in Python. Machine Learning Research, 12, 2825-2830.
[13] Randhawa K, Loo C K, Seera M, Lim C P, and Nandi A K. (2018). AdaBoost and majority voting are utilized for credit card fraud detection. IEEE Access, 6, 14277-14284.
[14] Rehmsmeier, M., and Saito, T. (2015). The precision-recall plot contains more information than the ROC plot for evaluating binary classifiers on unbalanced samples. 10(3): e0118432, PLoS ONE.
[15] Machine Learning Group, ULB. (2013). dataset for credit card fraud detection. Kaggle
[16] Codefinity: Courses with certificates | Online Learning Platform. (n.d.). Codefinity.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY