International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EFFECTS OF IMBALANCED DATA ON BANKING FRAUD DETECTION MODELS

Hung K. Nguyen[1], Tu T. Nguyen[*2], Oanh T. K. Nguyen[3], Ho Thi Hoang Luong[4]

[1]*Civil Engineering Department, Lac Hong University, Dong Nai, Vietnam.*

[2]*Civil Engineering Department, Lac Hong University, Dong Nai, Vietnam;*

[*]*Corresponding author email: tunt@lhu.edu.vn*

[3]*Marketing Department, Thuongmai University, Hanoi, Vietnam*

[4]*Nghe An University, Vietnam*

Abstract:

The quality of data is crucial for the effective performance of machine learning (ML) predictive models. To create a high-quality dataset, input data should be thoroughly processed to remove missing values, entry errors, outliers, and data imbalance. In this study, various techniques to address imbalanced datasets, namely, random under-sampling (RUS), random oversampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN), were tested using a highly imbalanced credit card transaction dataset. A total of 284807 credit card transactions were downloaded from previously published sources, with 28315 legitimate records and 492 frauds identified. In the first phase, the performances of four commonly used ML predictive models, Logistic Regression (LR), Random Forest (RF), Gaussian Naive Bayes (GNB), and Extra Trees (ET), were evaluated using a balanced subset of 984 records to identify the best-performing model. In the second phase, the selected ML model was used to assess the effectiveness of each of the data-modification techniques using three metrics: *Precision, Recall*, and F_{1score}. Results showed that the ET model performed best with a dataset modified using the SMOTE technique, achieving *Precision, Recall,* and F_{1score} values of 0.8787, 0.8613, and 0.8699, respectively. In contrast, the dataset modified with RUS yielded poorer results, with a *Precision, Recall*, and F_{1score} of 0.0788, 0.9208, and 0.1452, respectively. Finally, an important analysis was conducted, and the ET model configuration utilizing eight key features was recommended for this dataset.

Keywords: Fraud Detection; Imbalanced Data; Machine Learning; Under-Sampling; SMOTE; Extra Trees Classifier; Features Importance.

1. Introduction

In the current technological era, businesses face a constant threat of financial fraud through various activities such as online transactions and financial services. This not only leads to monetary loss but also adversely affects customer satisfaction, making it a major challenge for companies to deal with. Therefore, identifying and preventing fraudulent activities is a must for businesses. While the traditional approach of rule-based detection has been widely used, it

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

is time-consuming and requires manual updating of rules. However, with the advent of ML, a new approach to security systems has emerged, which enables faster and more efficient identification of hidden patterns and correlations. This method can provide accurate and timely detection of suspicious activities, thus reducing the financial losses incurred by businesses due to fraudulent activities.

To develop an effective ML model, the model first needs to be carefully trained using historical records so that the model can classify between legitimate and fraudulent activities. In the credit card operations domain, over 99% of the transactions are legitimate, and only a small amount of the transactions (say, less than 1%) are fraudulent, making credit card transaction data highly imbalanced. This could pose a challenge for training ML models, as the models may tend to focus on the majority class (legitimate transaction) and not learn the characteristics of the minority class (fraudulent) due to their limited occurrence [1]. This issue appears not only in fraud prevention areas [2-4], but also in software development [5], disease diagnostics [6], and other fields [7-11].

Imbalanced input data could result in poor performance, such as low accuracy for the minority class, and cause a negative impact on the accuracy of predictive ML models. Since the models may either not accurately capture the characteristics of the minority class, present a low precision for the minority class, or may fit too closely to the majority class, leading to poor generalization to unseen data. Many techniques have been developed for handling imbalanced data, such as resampling, ensemble methods, cost-sensitive learning, and threshold adjustment [12]. Within the context of this research, four variations of the resampling approach, including RUS, ROS, SMOTE, and TL, were employed to investigate the effectiveness of each technique on the performance of a selected ML model. Figure 1 shows the two sub-categories of the resampling technique.

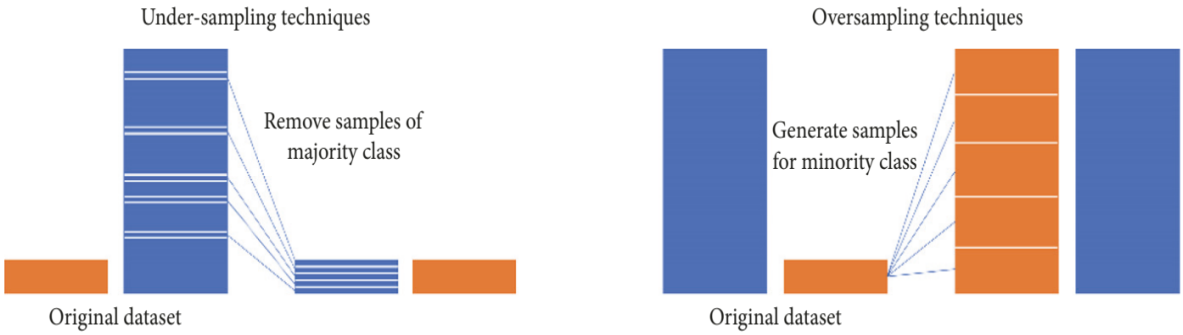


Figure 1. Illustration of oversampling and under-sampling techniques

Random under-sampling is a method for handling class imbalance in ML by reducing the size of the majority class in the training data [13]. It is undertaken by randomly picking a subset of the majority class data and removing instances in the majority until the class distribution is balanced. This approach is rapid and simple, but can result in the loss of valuable information and patterns in the majority of class data [14]. Furthermore, it can lead to over-

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

fitting if the reduced data set is too small, as the model may be too closely fit to the limited data. A different variation of RUS has been proposed to improve the quality of modified data. For example, Mani [15] used the NearMiss approach to select the records that are nearest to the minority samples. Lemaitre [16] applied Cluster Centroid to replace a cluster of majority records with the cluster centroid of K-means algorithms.

By contrast, the ROS approach, proposed by Batista et al. [17], works by simply duplicating instances of the minority class until the class distribution is balanced. This method avoids the loss of information from the majority class that can occur with under-sampling, but can result in over-fitting and may also lead to a more computationally expensive training process. In 2002, a more advanced oversampling method called SMOTE was proposed by Chawla et al. [18]. Instead of replicating the existing samples as in ROS, SMOTE works by generating synthetic samples of the minority class based on their feature space neighbors to balance the class distribution. New synthetic samples are created by linearly interpolating the features of a randomly selected minority class instance and one of its nearest neighbors, as illustrated in Figure 2. SMOTE generates new samples that are similar to existing minority class instances, but with slight variations, effectively increasing the size of the minority class. This technique can help the model learn more effectively from the minority class, without losing information from the majority class as with under-sampling or risking over-fitting as with over-sampling.

Another variant of SMOTE to investigate is ADASYN, which was first introduced by He et al. [19]. This algorithm used a weighted distribution for different minority class records depending on the level of difficulty in learning; thus, it can reduce the bias introduced by the class imbalance and adaptively shift the classification decision boundary toward the difficult records. The major difference between the SMOTE and ADASYN techniques is that they identify the minority class examples in regions where the majority class dominates. This enables ADASYN to create synthetic samples in the less dense regions of the minority class.

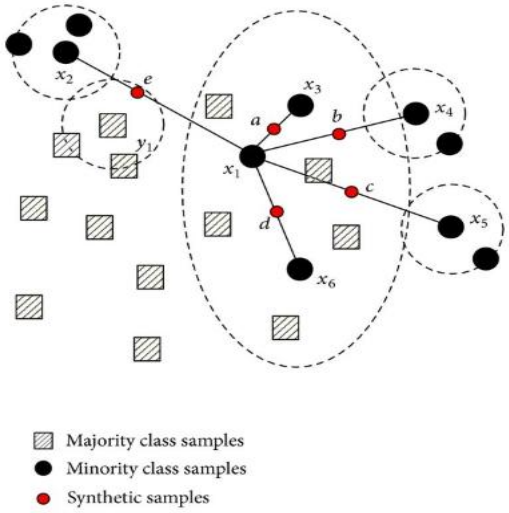


Figure 2. Illustration of the Synthetic Minority Over-Sampling Technique

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In this study, different data manipulation techniques, namely RUS, ROS, SMOTE, and ADASYN, were employed to investigate the effects of data manipulation on the performance of ML models. Highly imbalanced credit card transaction data downloaded from the previously published source [20] were utilized. In the first phase, the records from the minority class (fraud samples) in the original dataset were selected, and then the under-sampling technique was applied for the majority class (non-fraud records) to create a new subset of data with an equal number of records between the two classes. The subset data were then employed to choose the best-performing ML model among various commonly used models (LR, RF, GNB, and ET) based on three metrics, including Precision, Recall, and F_{1score}.

In the second phase, the selected ML model was then employed to investigate the effectiveness of each data manipulation technique. Additionally, the feature importance analysis was conducted to calculate the predictive power of the ML model for all input variables in the training dataset. Based on feature importance analysis results, the performance of different model configurations was evaluated, and model improvement was recommended. Python programming language and various libraries such as SciKit-Learn [21], Pandas [22], Matplotlib [23], and Seaborn [24] were employed to develop the coding for this study. The subsequent section provides details on the original and the subset data.

2. Data description

This previously published tabulated credit card transaction data contains 284,807 records with 28 principal features related to the details of the users [20]. The principal component analysis transformation has been applied to the data to anonymize features related to information about cardholders. Additionally, the time in seconds between transactions and the purchase amount (in units) is also provided. All records have been assigned as legitimate records (class 0) or fraudulent transactions (class 1). Table 1 shows some fundamental statistics of the selected features in the dataset.

Table 1. Statistics of the selected features

Features	mean	std	min	0.25	0.5	0.75	max
V1	94814	47488	0	54202	84692	139321	172792
V2	3.92E-15	1.959	-56.408	-0.920	0.018	1.316	2.455
V3	5.68E-16	1.651	-72.716	-0.599	0.065	0.804	22.058
V4	-8.8E-15	1.516	-48.326	-0.890	0.180	1.027	9.383
...	...	...	...	...	...	...	...
V25	4.46E-15	0.606	-2.837	-0.355	0.041	0.440	4.585
V26	1.43E-15	0.521	-10.295	-0.317	0.017	0.351	7.520
V27	1.7E-15	0.482	-2.605	-0.327	-0.052	0.241	3.517
V28	-3.7E-16	0.404	-22.566	-0.071	0.001	0.091	31.612

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2.1. Characteristics of the original dataset

Figure 3 shows the distribution of the original dataset. It is characterized by a high level of imbalance, with the majority of samples falling into the class 0 category (indicating non-fraudulent transactions) and a much smaller proportion of samples belonging to class 1 (indicating fraudulent transactions). This disparity is clearly illustrated by the count of each target class, which reveals that there are 284,315 class 0 samples and only 492 class 1 samples.

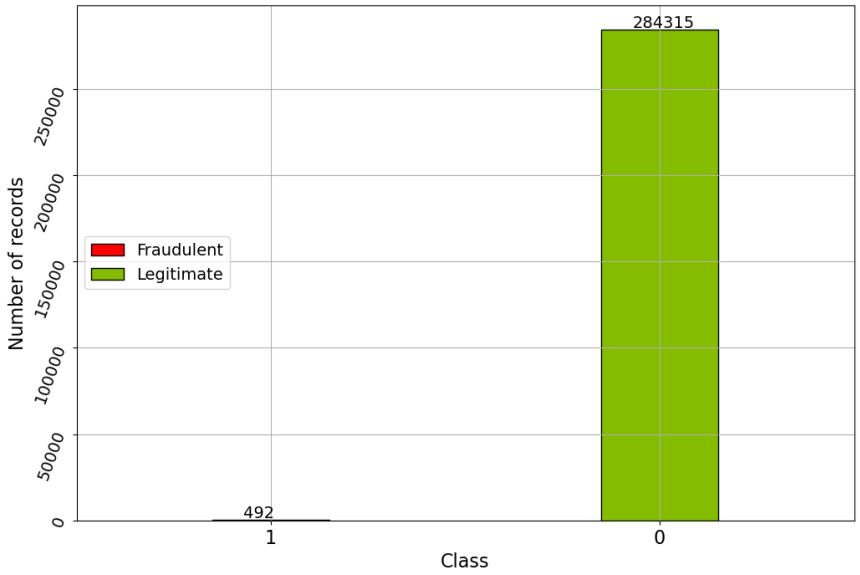


Figure 3. Class distribution of the original dataset

In order to visualize the distribution of each feature related to the information of the cardholders in the dataset, the histogram plots were performed for the 28 features, as shown in Figure 4. In each subplot, the height of the bar corresponds to the frequency of values of this bin, and the width of the subplot figure shows the range of the feature. As can be observed, the features V1, V4, V12, and V24 showed a high level of skewness, and most of the features presented a low level of outliers.

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

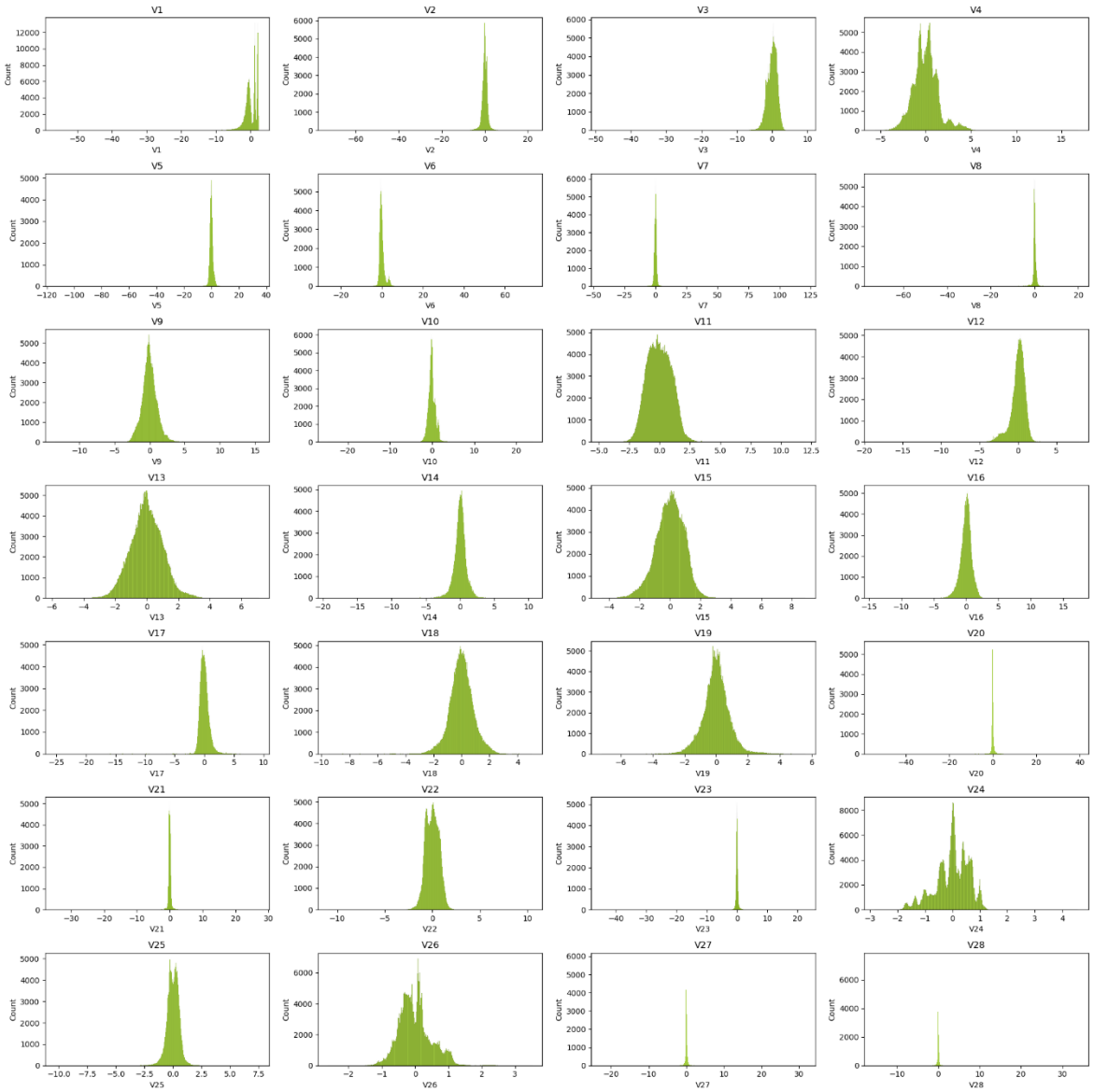


Figure 4. Histogram plots for inputs in the dataset

2.2. Sub-dataset for the selection of the predicted ML model

The Tomek Link (TL) technique was used in this study to create a balanced subset of data from the original imbalanced dataset. This technique handles class imbalance by removing instances of the majority class that are closely surrounded by instances of the minority class. Tomek Links are defined as pairs of instances, one from the minority class and one from the majority class, that are the closest neighbors to each other. The majority class instance is removed, as it is considered to be a noisy instance that is likely to interfere with the learning process. The newly created subset data consisted of an even distribution of fraud and non-fraud transactions. That means a total of 492 fraud records in the original dataset were employed, incorporating another 492 transactions selected from the 284315 legitimate records using the TL technique. Figure 5 shows the class distribution of the balance sub-dataset. The purpose of creating the sub-dataset is to test the performance of the four mentioned ML models as well as

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

to see the true correlation between the class and all features in the dataset, as presented in Figure 6.

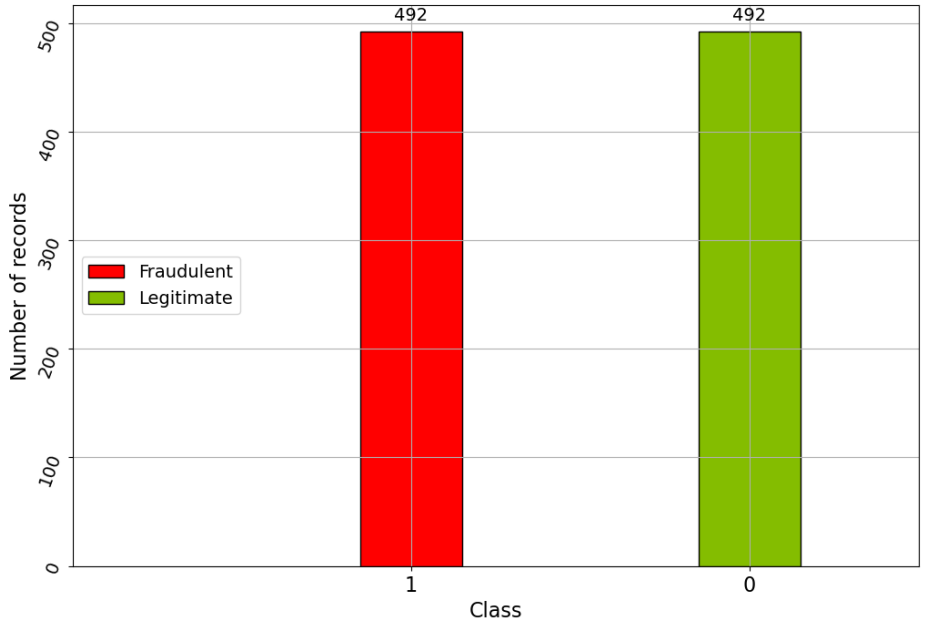


Figure 5. Class distribution of the sub-dataset

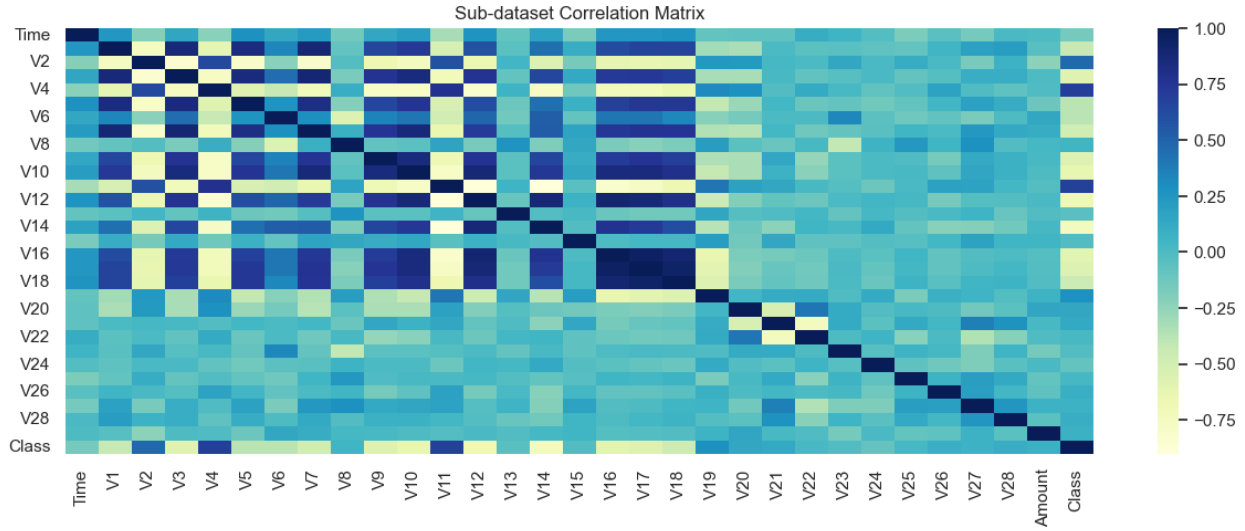


Figure 6. Correlation matrix between inputs and outputs

3. Research methodology

In this section, the three different types of model evaluation metrics were briefly described firstly, followed by the evaluation of the performance of the most popular ML classifiers with equally distributed input data to select the best ML model. Finally, the selected ML model was then employed to investigate the efficacy of four different data manipulation techniques, namely RUS, ROS, SMOTE, and ADASYN. Details are provided in the subsequent sections.

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.1. Performance metrics

Accuracy refers to the ratio of the number of correctly predicted types of transactions (i.e., fraud or legitimate) to the total number of transactions. However, for the classification problems using a highly imbalanced dataset in this study, the accuracy indicator is not a good measure since it gives equal importance to both false positives and false negatives; instead, other performance metrics, including Precision, Recall, and F_{1score} were used. Precision, on the one hand, can be understood as out of all transactions that were predicted as fraudulent, how many truly fraudulent transactions are? The recall, on the other hand, can be identified as out of all the fraud transactions, how many of them were predicted correctly. F_{1score} is the harmonic mean of precision and recall. Precision, Recall, and F_{1score} can be calculated through equations (1a-1c) using true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) variables.

$$Precision = \frac{TP}{TP + FP} \tag{1a}$$

$$Recall = \frac{TP}{TP + FN} \tag{1b}$$

$$F_{1score} = 2 \times \frac{Precision \times Rcall}{Precision + Recall} \tag{1c}$$

An alternative way to present the performance results is to use a confusion matrix, as illustrated in Figure 7. The columns of a confusion matrix represent the true value, and the rows show the predicted values assigned by the predictive model. The element a_{ij} (i is the row, and j is the column) indicates that the model assigned the value as i while the true value in the database is j. The elements in the diagonal of the confusion matrix (a_{ii} in the light green cells) are the components correctly classified by the model.

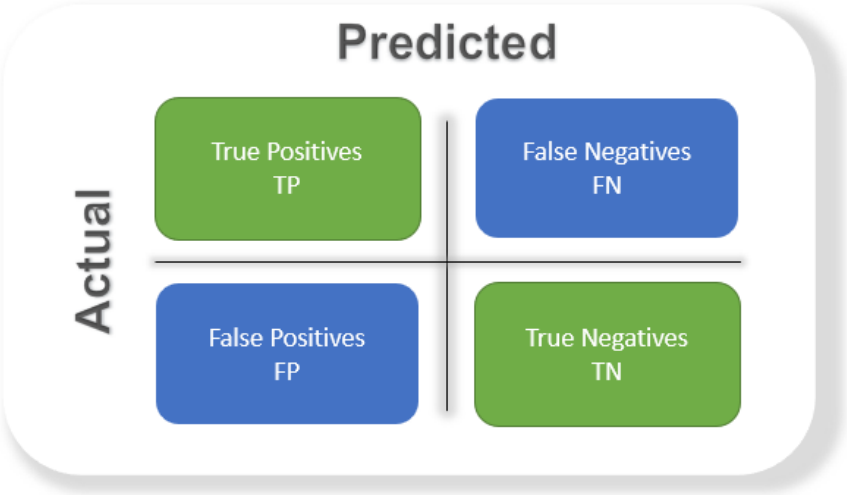


Figure 7. The schematic diagram of a confusion matrix

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2. Selection of classifiers for study

The balanced dataset created using the TL technique with a total of 984 samples was then utilized to investigate the performance of the five most common predictive ML algorithms, namely: LR, RF, GNB, and ET. To obtain a reliable result, the K-fold cross-validation method was employed, and the average results were calculated. Figure 8 presents the overall performance comparison of the above-mentioned algorithms in terms of Precision, Recall, and F_{1score}.

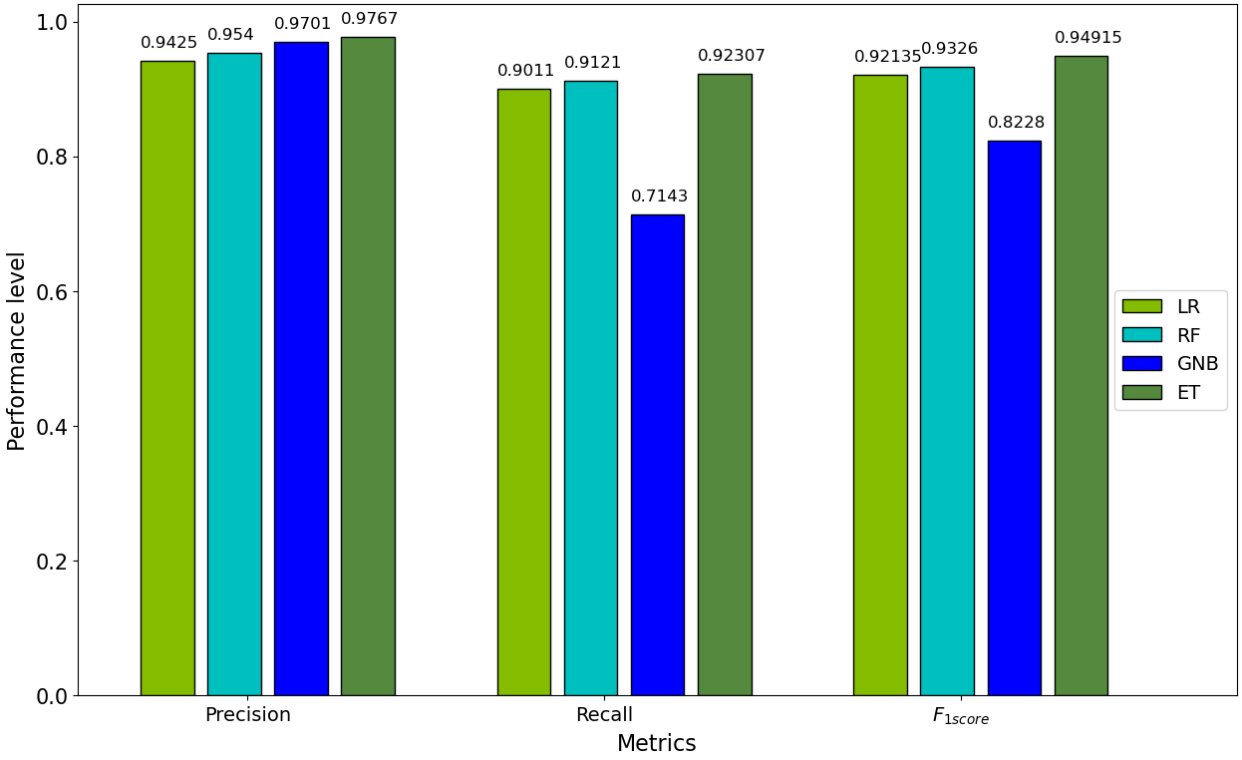
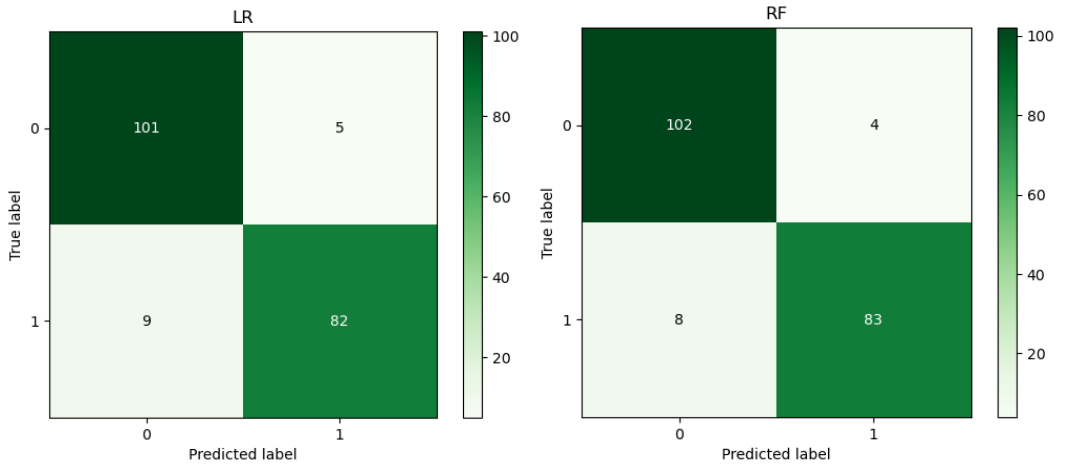


Figure 8. Performance of the investigated algorithms

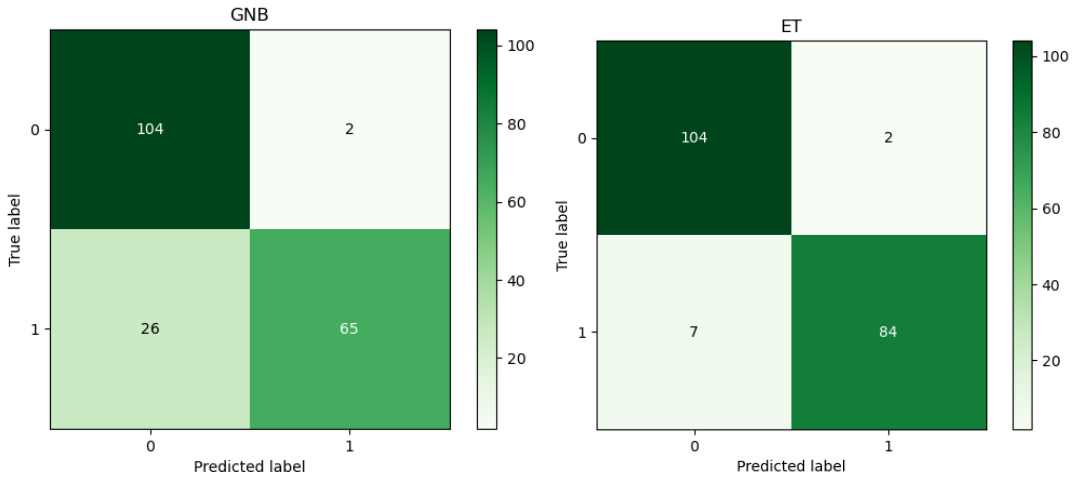
Figure 9 presents confusion matrices of four algorithms for the testing dataset. As can be seen clearly, the ET classifier showed the best performance for the sub-dataset with higher values in all three metrics compared to the other classifiers. For that reason, the ET was selected in this study to perform further investigation with a modified dataset.

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



(a) Logistic Regression (b) Random Forest



(a) Gaussian Naive Bayes (b) Extra Tree

Figure 9. Confusion matrix for the testing dataset of different classifiers

3.3. Handling imbalanced input data

The original imbalanced dataset was first randomly divided into training and testing datasets with a ratio of 80/20. That means, 80% of the original dataset was used for training purposes, and the remainder 20% of the entire dataset was employed for testing the performance of the model. While the testing dataset was kept unchanged and imbalanced during the investigation, the training dataset was respectively manipulated using one of the four data manipulation techniques, namely RUS, ROS, SMOTE, and ADASYN. For each modified training dataset, the ML model was first trained and then tested with the original unbalanced data test set. Figure 10 illustrates all the steps applied to evaluating the efficacy of methods to handle imbalanced data.

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

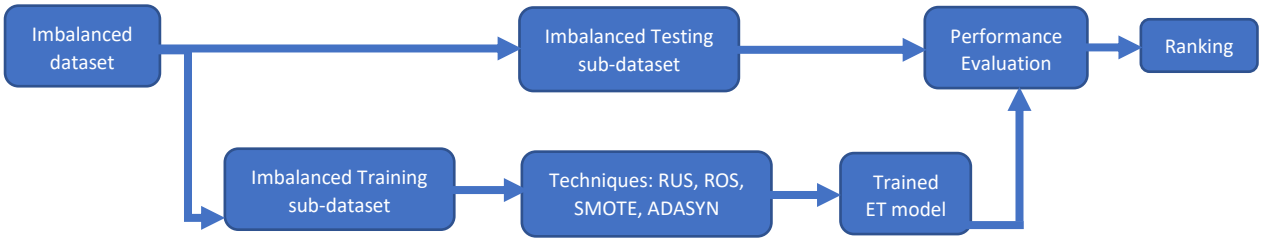


Figure 10. Flowchart of efficacy evaluation with different data manipulation techniques

4. Results and discussion

As previously mentioned, four data manipulation algorithms, namely RUS, ROS, SMOTE, and ADASYN, were used to modify the dataset for training the ET model. The performance of each trained ET model was then evaluated for the imbalanced test set data using Precision, Recall, and F_{1score} indicators. Finally, the importance analysis was implemented, and model improvement was investigated. Details of the work mentioned are presented in the following sections.

4.1 Performance of the Extra Trees algorithm

Figure 11 shows the performance of the selected ML algorithm using four data manipulation techniques. The performance of the ML algorithm was evaluated based on Precision, Recall, and *F1-score* indicators. As can be seen, the ET al algorithm performed well for the training dataset using the SMOTE data manipulation approach with a high value of all considered metrics as expected. By contrast, the performance of the ET model using a modified training dataset generated by the RUS technique showed a low level of generalization, with a value of Precision and F_{1score} were of 0.0788 and 0.1452, respectively. It is interesting to note that the value of the true positive rate (i.e., Recall) indicator, in this case, was high compared to other cases. It can be explained that the model was overfit for the majority class due to the lack of samples in the training dataset. Additionally, the difference in the model performance of the two trained ET classifiers using SMOTE and ADASYN data manipulation techniques was minimal. A possible explanation for this might be due to the original dataset being scattered enough. As a result, the synthetic samples that ADASYN created do not make a significant difference compared to those of the SMOTE technique.

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

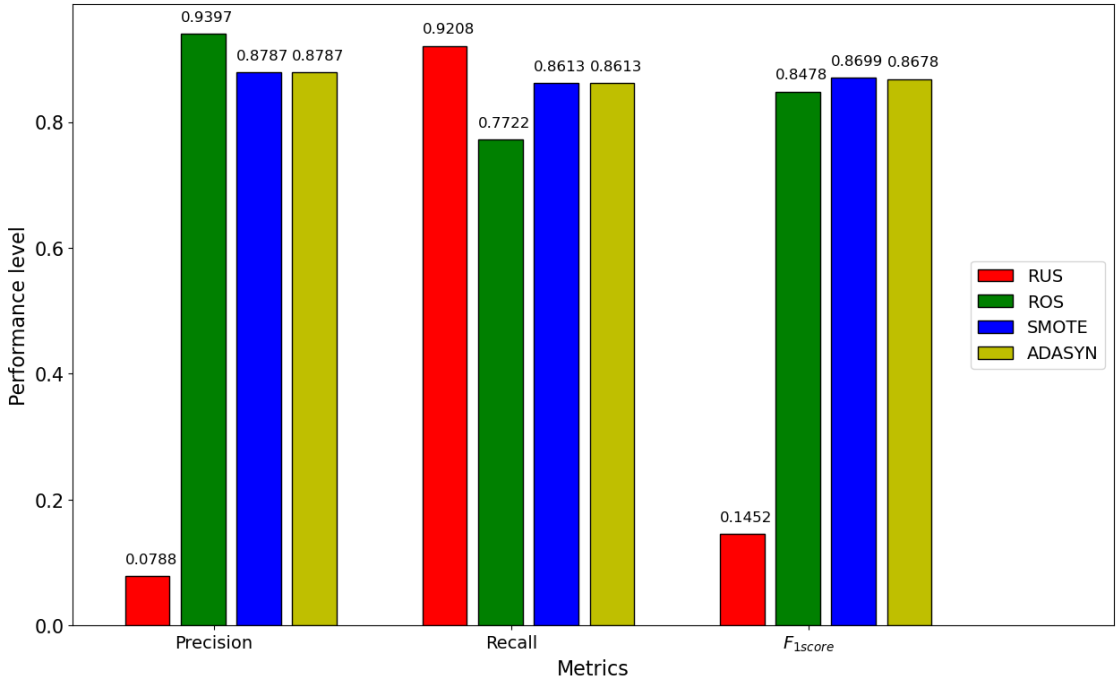


Figure 11. Performance of the ET classifier using different data manipulation techniques

With regard to the performance of the ET classifier using a modified training dataset applied with SMOTE, the results revealed an excellent classification capability of the trained ML model. The values of both Precision and Recall were well above 87%, showing that the model was reliable in classifying the correct samples in both classes. It is crucial since, as previously mentioned, the original data, as well as the test data, are highly imbalanced; the trained ML model could correctly categorize the records in the minority class (i.e., fraudulent records). Fraudulent transactions, in this case, seem more important than legitimate records since the occurrence of fraudulent transactions may cause substantial financial damage to the business.

4.2 Features importance

In order to understand the importance of each input parameter to the output of the ET classifier, feature importance was performed to calculate a score for all inputs of the dataset. An input variable with a higher score means that the input attributes a significant predictive power to the model. Figure 12 illustrates the analysis results of the trained ET model using the SMOTE technique to manipulate the training dataset. As can be seen, V4, V12, and V14 are the top three important parameters for the ET model among all variables in the credit card transaction dataset. Other variables, such as time, V13, V15, V22, V23, V24, V25, V26, and transaction amount, are less important to the classifier.

Received: August 15, 2025

International Journal of Applied Mathematics
Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

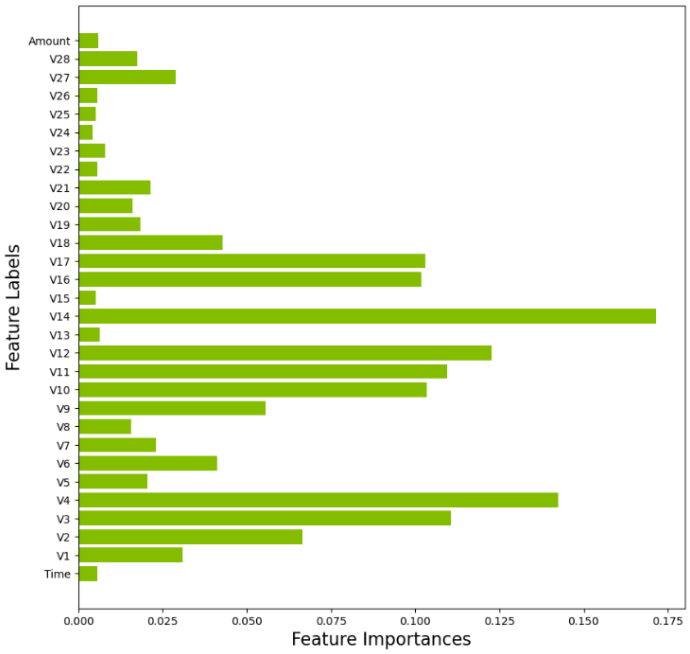


Figure 12. Features Important analysis

4.3 Model recommendation

Different configurations of the ET classifier were developed by changing the number of estimators to investigate the effects of model complexity on its performance. Based on the feature importance analysis results in Figure 12, three groups of estimators were selected depending on their importance score including (i) extremely importance (i.e., V4, V12, and V14), (ii) importance (i.e., V3, V4, V10, V11, V12, V14, V16, and V17), and (iii) least importance (i.e., Time, V13, V15, V22, V23, V24, V25, V26, and Amount). Table 2 lists the detailed performance of the various ET configurations using estimators in three groups. It is noted that the ET1 model was created by using all estimators in the original dataset, and it was used as the base model for comparison.

Table 2. Performance of the ET classifier with different configurations.

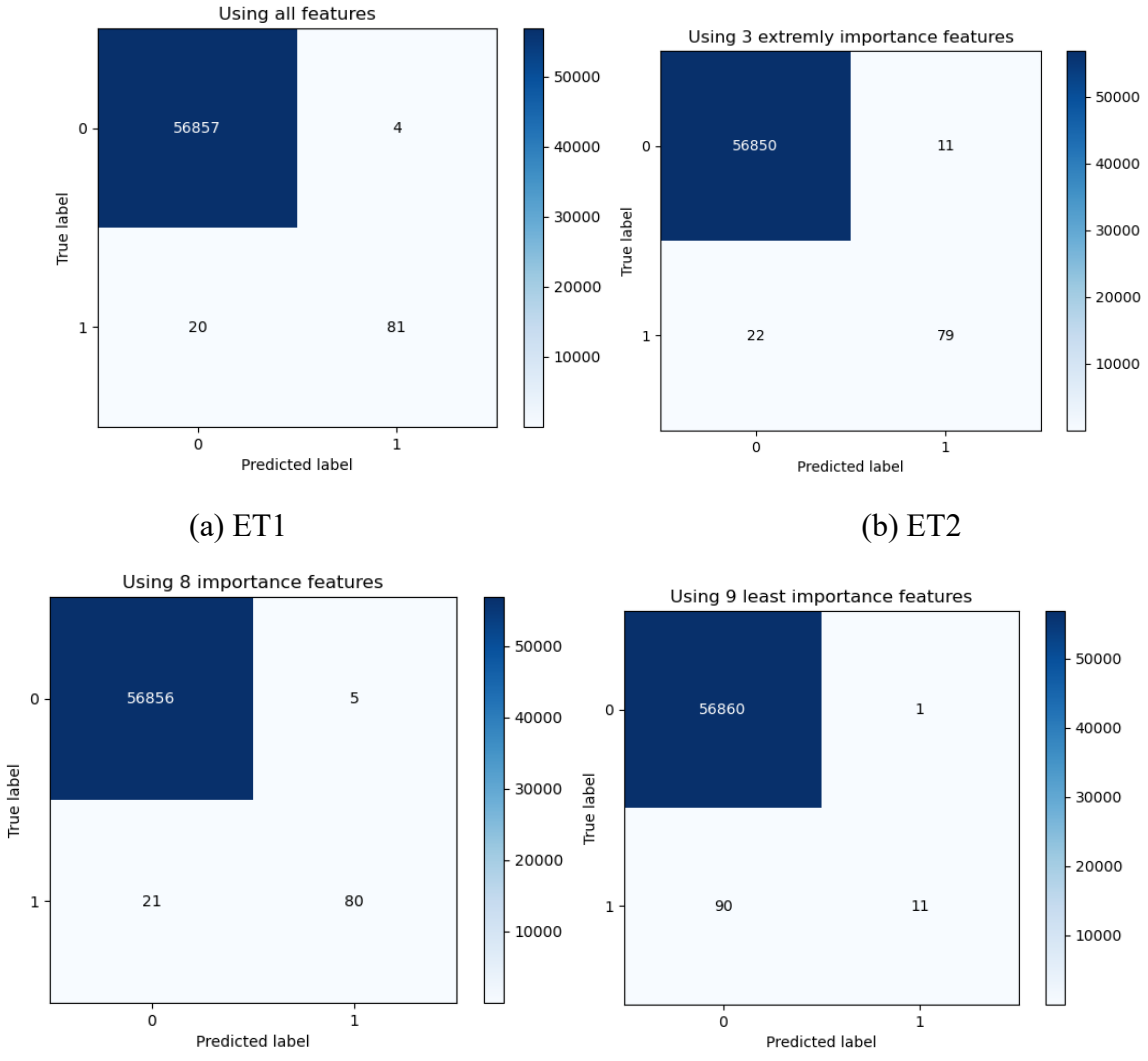
Model	Variables	Precision	Diff.	Recall	Diff.	F_{1score}	Diff.
ET1	All	0.9529		0.8019		0.8709	
ET2	Extremely important group	0.8876	-6.85%	0.78221	-2.46%	0.8316	-4.51%
ET3	Important group	0.9411	-1.24%	0.7920	-1.23%	0.8602	-1.23%
ET4	Least important group	0.9207	-3.38%	0.1188	-85.2%	0.2105	-75.8%

An alternative way to visualize the performance of the ET model using different configurations is to utilize confusion matrix plots, as illustrated in Figure 13. As can be seen, for the ET2 configuration, only three important features out of the total 31 were used, and the performance of the ET model showed only a minor drop in the performance metrics (i.e., a difference of 6.85% in Precision, 2.46% in Recall, and 4.51% in F_{1score} in comparison with the

Received: August 15, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

MT1 model). That means the predictive power of these three estimators was dominated over the other ones. By contrast, in the MT4 configuration, although many estimators were used, the predictive power of these variables was very low. Thus, the performance of the ET4 model was very poor, especially in Recall and $F_{1score,}$ with a difference compared to the base model was 85.2% and 75.8%, respectively, as in Figure 13d. Besides, the ET3 model employed 8 high predictive power estimators in the important group, which could result in the approximate classification capacity for the ET1 model using all estimators in the original training dataset. Thus, this less complex model (ET3) could be used for this dataset without losing predictive power (say, less than 2%).



(a) ET1 (b) ET2

(c) ET3 (d) ET4

Figure 13. Confusion matrix of different model configurations for the testing dataset

5. Conclusions

In this paper, the effects of different data manipulation techniques on the performance of the ET classifier were investigated. The original imbalanced dataset was modified using RUS, ROS, SMOTE, and ADASYN techniques to obtain the evenly distributed training

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

dataset. The performance of the ET classifier using different training datasets was evaluated and ranked. Results from the analysis revealed that the trained ET model using a training dataset developed by the SMOTE data modification technique provided the best performance with a Precision, Recall, and F_{1score} of 0.8787, 0.8613, and 0.8699, respectively. By contrast, trained ET models using a training dataset created with the RUS data manipulation method experienced the lowest classification capacity. Performance for the two trained ET classifiers using SMOTE and ADASYN data manipulation methods was found to be an insignificant difference.

With regard to the important features analysis results, V4, V12, and V14 were expressed as important parameters for the ET model, while other features, such as Time, V13, V15, V22, V23, V24, V25, V26, and Amoun, showed a limited predictive power for the ET model. Regarding the model complexity, within the context of this dataset, the trained ET model using 8 high predictive power estimators namely V3, V4, V10, V11, V12, V14, V16, and V17 was recommended for the dataset with a decrease in the accuracy of the predictive ML model of less than 2% compared to the base model with all 30 estimators. This could also reduce the burden of computation as well as the complexity of the model.

Data Availability

The data used to support the findings of this study are included in the article cited.

Conflicts of Interest

The authors hereby declare no potential conflicts of interest regarding the research, authorship, and/or publication of this article.

References

[1] Thabtah F, Hammoud S, Kamalov F, Gonsalves A. 2020. Data imbalance in classification: experimental evaluation. Information Sciences 513:429–441.

[2] Somasundaram A, Reddy S. 2019. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. Neural Computing and Applications 31(1):3–14.

[3] Zoričák M, Gnip P, Drotár P, Gazda V. 2020. Bankruptcy prediction for small-and medium-sized companies using severely imbalanced datasets. Economic Modelling 84:165–176.

[4] Le T, Vo MT, Vo B, Lee MY, Baik SW. 2019. A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction. Complexity 2019: Article 8460934.

[5] Yang X, Lo D, Huang Q, Xia X, Sun J. 2016. Automated identification of high-impact bug reports leveraging imbalanced learning strategies. In: 2016 IEEE 40th annual computer software and applications conference (COMPSAC), Vol. 1. Piscataway: IEEE, 227–232.

[6] Wang L, Zhao Z, Luo Y, Yu H, Wu S, Ren X, Zheng C, Huang X. 2020. Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods. Computer Methods and Programs in Biomedicine 196:105567.

Received: August 15, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[7] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. 2017. Learning from class-imbalanced data: a review of methods and applications. Expert Systems with Applications 73:220–239.
[8] Nguyen, T. T., & Dinh, K. (2019). Prediction of bridge deck condition rating based on artificial neural networks. *Journal of Science and Technology in Civil Engineering (STCE) - NUCE*, *13*(3), 15-25.
[9] Nguyen, T. T., Ngoc, L. T., Vu, H. H., & Thanh, T. P. (2021). Machine learning-based model for predicting concrete compressive strength. International Journal of Geomate, 20(77), 197-204.
[10] Nguyen, T.T., Pham, D. H., Pham, T.T., and Vu, H.H. (2020). Compressive strength evaluation of Fiber-Reinforced High Strength Self-Compacting Concrete with artificial intelligence. *Advances in Civil Engineering,* 2020, 3012139.
[11] Hung Cao, Hieu V. Phan, and Sabatino Silveri. Data Breach Disclosures and Stock Price Crash Risk: Evidence from Data Breach Notification Laws. Manuscript submitted for publication, 2020.
[12] He H, Garcia EA. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9):1263–1284.
[13] Li Y, Chai Y, Hu Y, et al. 2019. Review of imbalanced data classification methods. J. Control and Decision, 34(04):673-688.
[14] Liu, X.-Y.; Wu, J.; and Zhou, Z.-H. 2009. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39(2):539–550.
[15] Mani, I., and Zhang, I. 2003. knn approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets, volume 126.
[16] Lemaˆıtre, G.; Nogueira, F.; and Aridas, C. K. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research 18(17):1–5.
[17] Batista GE, Prati RC, Monard MC, 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 6(1):20–29.
[18] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357.
[19] He H, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. 1322-1328.
[20] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015
[21] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Received: August 15, 2025

Volume 38 No. 10s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[22] McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).
[23] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.
[24] Waskom, M., Botvinnik, Olga, O'Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, Qalieh, Adel. (2017). mwaskom/seaborn: v0.8.1 (September 2017). Zenodo. https://doi.org/10.5281/zenodo.883859.

Received: August 15, 2025