

# A credit card fraud detection approach based on ensemble machine learning classifier with hybrid data sampling

Khanda Hassan Ahmed<sup>a,b,\*</sup>, Stefan Axelsson<sup>a</sup>, Yuhong Li<sup>a</sup>, Ali Makki Sagheer<sup>c</sup>

<sup>a</sup> DSV, Stockholm University, Stockholm, Sweden

<sup>b</sup> Computer Science Department, College of Science, University of Sulaimani, Sulaymaniyah, Iraq

<sup>c</sup> College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq

## ARTICLE INFO

### Keywords:

Ensemble model  
Machine learning  
Data imbalance  
Credit card fraud detection

## ABSTRACT

The existing fraud detection methods present limitations such as imbalanced data, incorrect identification of fraudulent cases, limited applicability to different scenarios, and difficulties processing data in real-time. This paper proposes an ensemble machine-learning model for detecting fraud in credit card transactions. It also integrates the Synthetic Minority Oversampling Technique (SMOTE) with Edited Nearest Neighbor (ENN) to address the problem of the imbalanced datasets. The experimental results show that our approach performs better than the existing methods. Therefore, it will establish an essential framework for the ongoing investigations in developing more robust and flexible systems for fraud detection.

## 1. Introduction

There has been an increase in fraudulent activity in the financial sector, particularly in credit cards (Bagga et al., 2020). Fraud is the unauthorized use of someone's credit card by an individual for personal transactions without the permission or consent of the cardholder. The convenience of using credit cards for financial transactions increases the difficulty of identifying fraudulent behavior (Shah & Passi, 2021). The British Ministry of Finance Annual Fraud Reporting reveals that in 2022, a total of 1.2 billion pounds was unlawfully obtained via both approved and illegal criminal operations, resulting in a significant loss of 2300 pounds each minute. Significantly, 78 % of incidents involving Authorized Push Payment (APP) fraud were initiated using Internet channels, while 18 % took place through telephonic channels (Annual Reports, 2024). Remote purchase fraud is a common cause of financial losses. In this kind of fraud, criminals use stolen card information to transact online or via phone/mail. This led to very huge losses (Khalid et al., 2024).

Due to the rising prevalence of credit card fraud, academics have been more interested in applying traditional Machine Learning methods and newer AI-based algorithms for fraud detection. However, the uneven data distribution is a significant obstacle to resolving the classification problem of fraudulent and genuine transactions. Given the

apparent disparity between the number of valid and fraudulent transactions, training a learning algorithm on the properties of the minority class becomes tough (Prasad et al., 2023; Sahithi et al., 2022).

The learning model would prioritize the majority class (non-fraud), which may inevitably lead to overfitting (Lokanan & Sharma, 2022). One challenge in Credit Card Fraud Detection (CCFD) is the infrequency of these fraudulent transactions compared with legitimate transactions. Hence, all recorded data will inevitably exhibit a significant disparity in the diversity of minority (fraud) and majority (legal) samples. To address the rising fraud associated with credit cards, it is necessary to design a highly accurate model that effectively meets the demands of credit card users.

Misclassification has consistently been a significant challenge in identifying digital credit card fraud in e-commerce systems (Barongo & Mbelwa, 2024). The study aims to minimize false positives to avoid inconveniencing clients and maintain faith in the financial system. By combining SMOTE and ENN, we proposed a method to enhance the dataset's balance and minimize the likelihood of overfitting to noisy or misleading instances. The proposed method can improve the generalization performance of the classifier, particularly on datasets with skewed distributions. The comparison between the results using our proposed Ensemble Machine Learning method based on the Sample Balancing technique (EML-SB) and those using supervised learning

\* Corresponding author.

E-mail addresses: [khanda@dsv.su.se](mailto:khanda@dsv.su.se), [khanda.ahmed@univsul.edu.iq](mailto:khanda.ahmed@univsul.edu.iq) (K.H. Ahmed), [stefan.axelsson@dsv.su.se](mailto:stefan.axelsson@dsv.su.se) (S. Axelsson), [yuhongli@dsv.su.se](mailto:yuhongli@dsv.su.se) (Y. Li), [ali\\_makki@uoanbar.edu.iq](mailto:ali_makki@uoanbar.edu.iq) (A.M. Sagheer).

<https://doi.org/10.1016/j.mlwa.2025.100675>

Received 13 September 2024; Received in revised form 22 April 2025; Accepted 17 May 2025

Available online 18 May 2025

2666-8270/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

shows that our EML-SB can efficiently increase the accuracy in detecting fraud activities on credit cards.

This paper is structured as follows: First, a review of related literature is discussed, highlighting the distinctions between our approach and existing methods. Then, we describe how the dataset was collected and processed. Based on this foundation, a detailed description of our model is provided. Finally, we outline the evaluation criteria, explain our experiments, and conduct a comparative analysis of our results with those of prior studies.

## 2. Related work

Wahab et al. in 2024 compared the effectiveness of a Deep Learning (DL) model to other machine learning models, including Adaboost and Decision Tree (DT) (Wahab et al., 2024). The study aims to determine whether particular DL parameters are responsible for the noted credit card default prediction accuracy improvements. The UCI ML repository is used in this study to retrieve the credit card defaulted customer dataset. Exploratory Data Analysis (EDA) is then used to visually show the results after various preprocessing procedures have been applied to the unprocessed data.

Mim et al. in 2024 proposed an approach to detect credit card fraud on unbalanced data using soft voting ensemble learning (Azim Mim et al., 2024). Several sophisticated sampling strategies (such as hybrid sampling, under-sampling, and over-sampling) are compared and evaluated with the suggested method to address the issue of class imbalance. Develop a variety of credit card fraud classifiers, both with and without sampling strategies, including ensemble classifiers. The experimental findings show that the suggested soft-voting method performs better than individual classifiers.

Tekkali & Natarajan in 2024 compared three well-proven and highly successful optimization methods: Root Mean Squared Propagation (RMSprop), Adaptive Moment Estimation (ADAM), and Stochastic Gradient Descent (SGD). Credit card fraud detection (CCFD) uses these optimization strategies on the Deep Convolutional Neural Network (DeepConvNet). Conclude that all four methods are appropriate for our CCFD assignment after carefully examining the nature of the problem, the properties of the objective function, and the computing factors. Experiments show that RMSprop performs better than the others, with an astounding 99.93 % accuracy (Tekkali & Natarajan 2024).

Prasad et al. in 2023 provided an ensemble approach to enhance CCFD. The authors concentrate on enhancing model parameters, refining performance metrics, and using deep learning techniques to rectify identification mistakes and minimize instances of false negatives. This research significantly increases the efficiency of the credit card fraud detection system by merging several classifier ensembles and carefully evaluating their performance. However, the model's poor performance appears in the evaluation parameters (Prasad et al., 2023).

Sahithi et al. in 2022 presented an approach for CCFD. Their approach used a Weighted Average Ensemble technique to merge the predictions of logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Adaboost, and Bagging models. This study demonstrates that their combined algorithm can accurately identify instances of credit card fraud within this field (Sahithi et al., 2022).

Qaddoura et al. in 2022 conducted a study to assess the efficacy of several oversampling techniques. The study included a range of machine-learning algorithms. The researchers discovered that over-sampling techniques may enhance the model's performance, albeit the specific approach is contingent upon the machine learning algorithm employed. However, the approach's computing overhead may limit its application in real-world situations (Qaddoura & Biltawi, 2022).

Another investigation was employed by Tanouz et al. in 2022 on using machine learning techniques to categorize credit card fraud. The DT, RF, LR, and Naïve Bayes (NB) were assessed, specifically considering datasets with a skewed distribution. The comprehensive analysis demonstrates that Random Forest is very efficient in detecting credit card

fraud, a crucial aspect of ensuring financial security. This model's performance is hindered by the absence of feature selections (Tanouz et al., 2021).

Ruttala et al. in 2020 carried a comparative analysis of RF and AdaBoost in the domain of CCFD. The investigation revealed comparable levels of accuracy when comparing the two methods. The RF performed better in terms of various evaluation matrices than Adaboost. Nevertheless, the dataset was biased, and there is no explicit indication of how this problem was resolved (Sailusha et al., 2020).

Sadgali et al. in 2019 ascertained an efficient methodology for identifying fraud. Notably, this work conducted a thorough and critical examination of past academic research without a specific dataset to analyze. Their findings emphasized Naive Bayes' superior performance, which attained the highest accuracy of 99 %. The SVM achieved an accuracy of 98 %, closely followed by the genetic algorithm, with an accuracy of 95 % (Sadgali et al., 2019).

Raghavan et al. in 2019 sought to identify abnormalities or fraudulent activities via data mining methods. They used three separate datasets from Australia, Germany, and the European Union to accomplish this purpose. Their study used SVM, KNN, and RF. Additionally, they constructed two distinct ensembles. However, its accuracy was poor across all utilized datasets (Raghavan & Gayar, 2019).

Saputra et al. in 2019 conducted a comparative analysis of the efficacy of various machine-learning methodologies. SMOTE addressed the challenges of an unbalanced dataset from the Université Libre de Bruxelles (ULB) Machine Learning Group sourced from Kaggle. The analysis indicated that the Neural Network achieved the best accuracy at 96 %. However, while 96 % accuracy in Credit Card Fraud Detection is impressive, it still leaves room for a significant number of false positives and false negatives (Saputra & Suharjito, 2019).

Therefore, we observe the following challenges: minimizing false positives to avoid inconveniencing clients and maintaining faith in the financial system is crucial. Optimizing the trade-off between incorrect positive and negative results (undetected fraud instances) is a sensitive issue requiring precise adjustment of detection algorithms and threshold configurations.

Combining the SMOTE with the ENN can effectively address the challenge of balancing false predictions in fraud detection. SMOTE helps by generating synthetic samples for the minority class (fraud cases), enhancing the model's ability to detect fraudulent activities. Meanwhile, ENN refines this approach by removing ambiguous and potentially misclassified samples from both classes. By ensuring a more balanced dataset, this combination enhances the efficiency of the detection algorithm and reduces the possibility of false positives, maintaining system integrity and client trust.

## 3. Methodology

### 3.1. Data collection

The dataset used in this work is sourced from the ULB Machine Learning Group, and its description is available at Kaggle (Credit Card Fraud Detection, 2024). It provides information on the transactions that took place during two days, including a total of 284,807 individual transactions. The fraud instances, which belong to the minority class, account for 0.172 % of the transaction data. The dataset exhibits a significant imbalance and bias towards the positive class. The dataset contains continuous numerical input variables obtained by a Principal Component Analysis (PCA) feature selection process, resulting in 28 principal components. Therefore, the current study employs a total of 30 input features (characteristics). The 'class' feature is the target column for binary classification, with a value of 1 indicating a fraud case and 0 indicating a non-fraud case.

### 3.2. An ensemble machine learning classifier

The model described in this article is shown in Fig. 1. The methodology used in this study is an ensemble machine-learning technique that integrates many classifiers, each selected for its unique capabilities. The AdaBoost classifier is a meta-estimator that starts by training a classifier on the original dataset. It then iteratively trains more copies of the classifier on the same dataset, with the weights of misclassified instances modified to choose the most difficult instances in subsequent classifiers (Taha & Malebary, 2020). RF constructs resilient decision trees (Dileep et al., 2021) (Dileep et al., 2021), while KNN classifies data by determining the majority class among its closest neighbors (Ittoo et al., 2021). The Voting Classifier, which collects predictions from various classifiers, is a significant addition (Vairam et al., 2022). Each option was carefully selected based on their proven efficacy in the previous studies, as extensively outlined in the literature review. Including many classifiers in this ensemble is a strategic approach intended to enhance the predictive capability of the proposed model.

Ensemble approaches effectively deal with imbalances in class distribution, showing strong performance in finding minority classes. They enable the combination of several less accurate learners, improving the model's overall predicting abilities (Chhabra et al., 2024).

Each Adaboost, RF, and KNN have unique characteristics and advantages that contribute to their performance and efficiency. RF can handle high-dimensional data and is resilient to overfitting. Adaboost is excellent at managing weak classifiers and increasing overall accuracy. KNN performs well with small datasets and is straightforward to implement. These algorithms can be very effective for fraud detection in credit card transactions, which require a high degree of accuracy and flexibility in response to changing fraud patterns.

To effectively apply these techniques for credit card fraud detection, it is crucial to address key challenges, including ensuring proper data preparation, managing class imbalances, optimizing hyperparameters, and enhancing scalability. Moreover, no existing work has fully leveraged this combination of methods in a unified framework, highlighting the novelty of this approach.

### 3.3. SMOTE

A statistical method was employed to increase the number of instances belonging to the minority class in a dataset while maintaining a balanced distribution. The component generated novel instances based on the minority situations given as input. In the case of SMOTE, the minority class (fraud) (designated as 1) was oversampled to match the

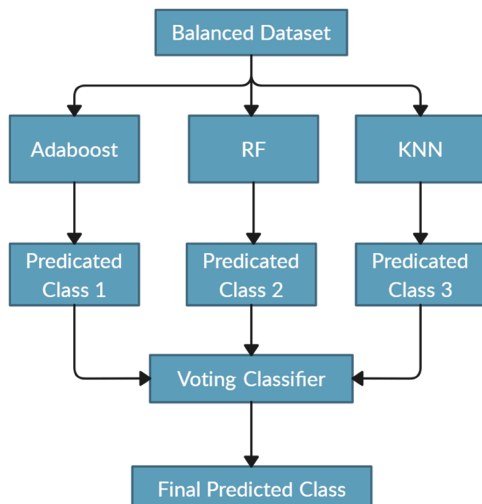


Fig. 1. The machine learning classifier based on (AdaBoost, RF, and KNN).

number of entries in the legitimate class. This was done to ensure that both classes had an equal number of entries for optimum training of models. In addition to under-sampling, both classes were combined to create a single dataset (Niveditha et al., 2019; Bounab et al., 2024).

### 3.4. Evaluation metrics

The examination and assessment of the performance metrics obtained during the evaluation provide a comprehensive analysis of the performance of each model. The following measurements were used to evaluate the proposed model.

Precision refers to the accuracy of the model's optimistic predictions (Yacoubly & Axman, 2020). Recall is the ratio of accurate optimistic predictions to the number of positive samples (both true and false negatives) (Yacoubly & Axman, 2020).

The F1 score is a quantitative measure that integrates the outcomes of accuracy and recall into a unified number (Fourure et al., 2021). Accuracy refers to the proportion of correct prediction entries in the samples (Heydarian et al., 2022).

### 3.5. The proposed credit card fraud detection approach

This section presents the proposed process. As shown in Fig. 2, the proposed approach begins with dataset selection, followed by data preprocessing, which includes data cleaning and normalization to ensure consistency and quality. After preprocessing, oversampling, undersampling, and a hybrid data sampling technique are applied to balance the dataset. The dataset is then split into training and testing subsets to develop and evaluate the model.

The ensemble model is constructed using Adaboost, KNN, RF, and a voting classifier, combining their strengths for improved fraud detection performance. The training phase involves feeding the prepared data into the selected models, after which the trained model is evaluated on the testing set. Finally, the system determines whether a transaction is fraudulent or non-fraudulent, with its effectiveness assessed through performance metrics and outcome analysis.

The methodology of the proposed model is described in the following Algorithm 1:

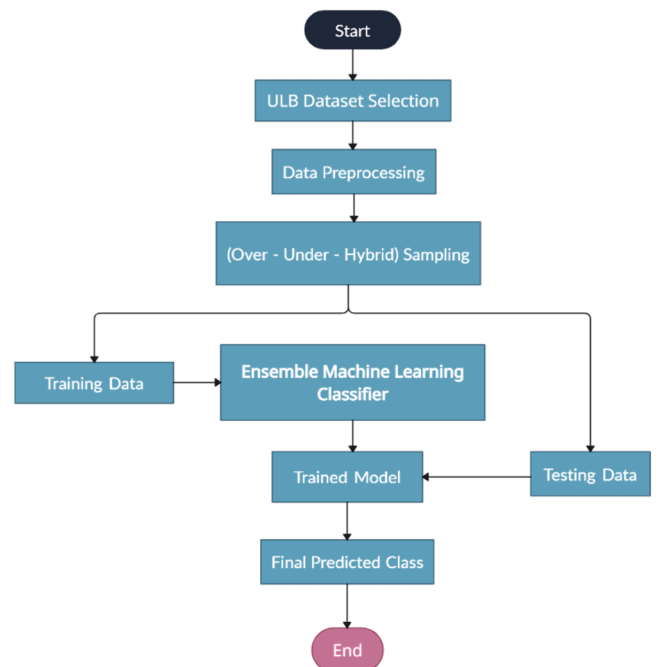


Fig. 2. The proposed ensemble machine learning with a hybrid data sampling technique.

**Algorithm 1**

The proposed credit card fraud detection.

**Input:** CCFD Dataset**Output:** Models confusion matrix

---

```

1. dataset = Load Dataset ()
2. preprocess data (dataset)
3. models [AdaBoost, RF, KNN, EML without sampling, EML with under-sampling, EML with smote, EML with (SMOTE + ENN)]
4. datasets [dataset without balancing, under-sampled dataset, SMOTE dataset, SMOTE + ENN dataset]
5. for  $m = 0$  to models count do:
6.   for  $d = 0$  to datasets count do:
7.     training data, testing data = Split (d)
8.     testing features = dropping (Class, d)
9.     testing labels = testing data [Class]
10.    confusion matrix (testing features, testing labels)
11.  end for
12. return (confusion matrix(m))
13. end for

```

---

**3.5.1. Data pre-processing**

The initial Proceeding is to preprocess the dataset to prepare it for implementation. During this stage, the data underwent processing using the following methods.

- Standardizing the 'Amount' column to facilitate analysis.
- Excluding the 'Time' column from the dataset since it had little impact on the training and assessment process.
- Verifying and eliminating duplicate items within the dataset.

The dataset was intrinsically limited due to the absence of information on its characteristics. Consequently, the feature selection and engineering strategies have been more attainable due to the need for distinct visibility of feature information.

**3.5.2. Under-Sampling**

From the majority class, a random sample was selected, consisting of legitimate transactions labelled as 0. The number of samples was chosen based on the needed ratio for the minority class. Within this part, to enhance the model's training, the number of entries for both classes was equalized by randomly selecting from the majority a sample size equal to that of the minority and combining the data into a single dataset.

**3.5.3. SMOTE and edited nearest neighbors**

ENN is a method used to reduce the size of a dataset by eliminating instances whose class label is different from the majority class label of its closest neighbors. The SMOTE algorithm is first used to create artificial samples for the minority class, augmenting its presence in the dataset. Afterward, the ENN algorithm processes the merged dataset (consisting of actual and synthetic samples) to eliminate noisy or misleading instances. Integrating SMOTE and ENN aims to efficiently tackle class imbalance while mitigating the risk of overfitting by removing noisy data.

To tackle the problem of imbalanced data, combining the SMOTE with the ENN can be a highly effective way to address the challenge of reducing and balancing false positives and negatives in CCFD.

**3.5.4. Model training**

After completing the sampling technique, data is divided into distinct training and testing sets, which is an essential stage in machine learning. The first stage involves randomly shuffling the dataset to guarantee that the training and testing sets accurately reflect the overall distribution. Next, 80 % of the shuffled dataset will be used for training purposes, and the remaining 20 % will be used for testing. This ratio guarantees that the model has enough data for learning while maintaining a distinct dataset for evaluating its performance. Next, the training set will train the model, enabling it to understand the connections between the input characteristics and the target variable. Assess the performance of the model using the testing set. This dataset serves as new and unseen data,

approximating the model's performance on data not included in the training dataset.

**4. Results and discussion**

Below is the performance evaluation for the proposed ensemble machine learning without sampling, with Under-sampling, SMOTE, and SMOTE + Tomek Links. the proposed ensemble machine learning model included KNN with 5 neighbors considered, RF with 100 decision trees, and Adaboost with 50 weak learners, followed by a voting classifier using soft voting.

**4.1. Ensemble machine learning without sampling**

**Table 1** shows the results acquired during the assessment of the proposed model (without sampling).

Based on the Confusion Matrix outcomes, the ensemble model (without sampling) and RF fared better than all other models, with 31 incorrectly predicted values. **Table 2** displays the accuracy values for each model in the testing samples.

The testing sample yielded findings indicating that the proposed ensemble model (without sampling) and the RF classifier achieved the maximum accuracy, with an ACC of 0.99943, surpassing all other models. The KNN classifiers had ACC values of 0.99941, which were carefully observed. The accuracy (ACC) of the AdaBoost classifiers was 0.99911. **Fig. 3** displays the graphical depiction of these results, illustrating how models react to the unseen data (testing data).

The ensemble model's F1 score, recall, and precision provide a more comprehensive evaluation of a classifier's performance than relying on overall accuracy (**Chicco & Jurman, 2020**). **Table 3** displays the outcomes for all models used on unseen samples (testing data) of the under-sampled data.

The results of the testing samples indicate that the suggested ensemble model (without sampling) and RF are the most effective models and achieved precision, recall, and an F1-score of 0.905, 0.736, and 0.812, respectively. By contrast, the AdaBoost method exhibits

**Table 1**

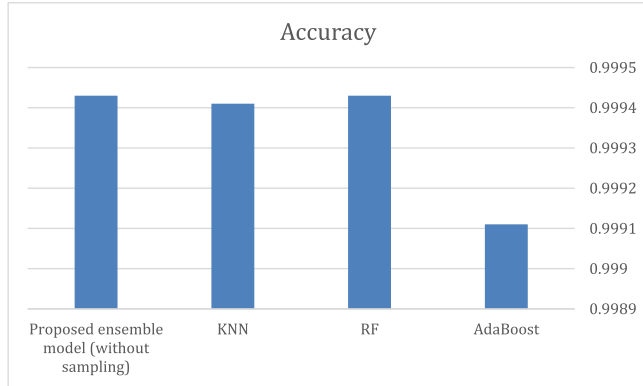
Confusion matrix values for the testing samples (without sampling).

Confusion Matrix	AdaBoost	RF	KNN	Proposed ensemble model (without sampling)
True Positives (TP)	63	68	68	67
True Negatives (TN)	55,021	55,034	55,033	55,035
False Positives (FP)	21	8	9	7
False Negatives (FN)	28	23	23	24

**Table 2**

Test the accuracy of the sample dataset (without sampling).

	AdaBoost	RF	KNN	Proposed ensemble model (without sampling)
Accuracy	0.99911	0.99943	0.99941	0.99943

**Fig. 3.** Accuracy results comparative analysis of all models on the testing data.**Table 3**

Evaluation matrix for the testing dataset (without sampling).

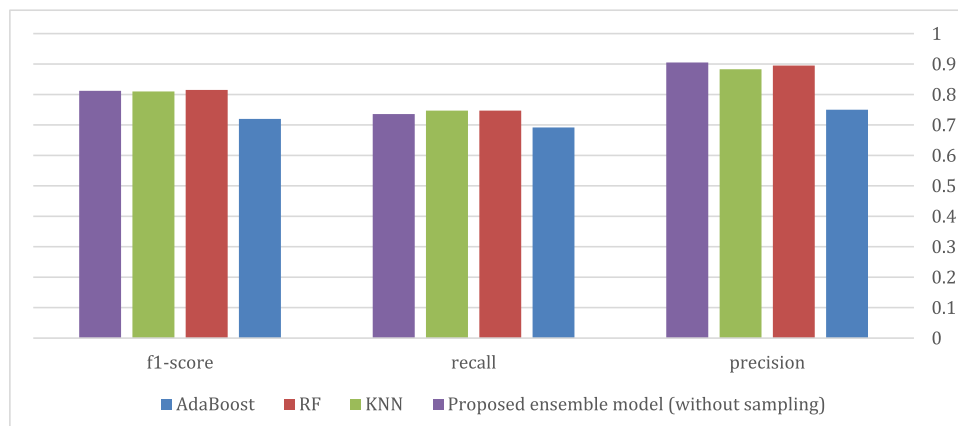
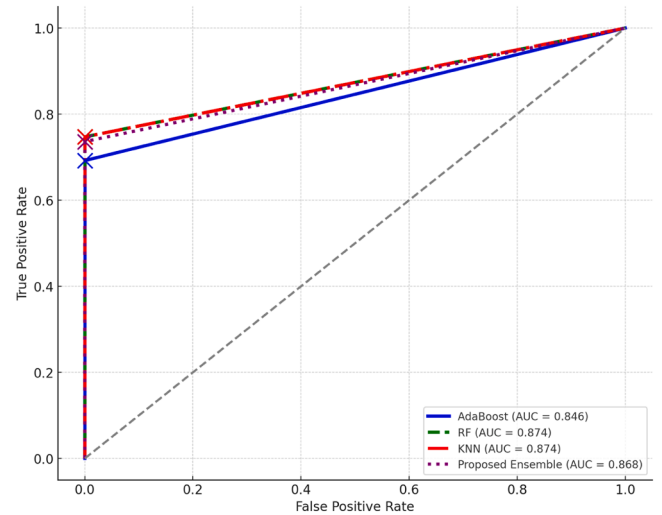
	AdaBoost	RF	KNN	Proposed ensemble model (without sampling)
precision	0.750	0.895	0.883	0.905
recall	0.692	0.747	0.747	0.736
f1-score	0.720	0.815	0.810	0.812

accuracy, recall, and F1-score values of 0.750, 0.692, and 0.720, respectively. The numerical values are shown in Fig. 4 to provide a more comprehensive comprehension of these findings.

Fig. 5 exhibits the Receiver Operating Characteristic (ROC) curve for all models, with each model's respective Area Under the Curve (AUC) values. It is worth mentioning that the ensemble model (without sampling), RF, and KNN showed the highest value for the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

#### 4.2. Ensemble machine learning with Under-Sampling

Table 4 displays the findings of the confusion matrix acquired when evaluating the suggested Ensemble Machine Learning model, which incorporates Under-Sampling. In our test, the strategy is set to 'auto'

**Fig. 4.** Evaluation matrix values of the ensemble model (without sampling) for the testing dataset.**Fig. 5.** ROC curve plots for the testing dataset.**Table 4**

Confusion Matrix values for the testing samples (with under-sampling).

Confusion Matrix	AdaBoost	RF	KNN	Proposed ensemble model (under-sampling)
True Positives (TP)	96	94	91	91
True Negatives (TN)	85	87	87	87
False Positives (FP)	3	1	1	1
False Negatives (FN)	6	8	11	11

(default), which means the majority class is under-sampled to match the size of the minority class (1:1 ratio).

Based on the data, the suggested ensemble model (with under-sampling) and KNN fared worse than all other models, with 12 incorrectly presented values. Table 5 displays the accuracy of all models on

**Table 5**

Test sample datasets accuracy results (with under-sampling).

	AdaBoost	RF	KNN	Proposed ensemble model (under-sampling)
Accuracy	0.95263	0.95263	0.93684	0.93684

the testing samples.

The ACC of the Proposed ensemble model (with under-sampling) and the KNN classifier in the testing sample data were the lowest among all models, achieving a value of 0.93684. The AdaBoost and RF classifiers had ACC values of 0.95263, which were quite close to one another. Fig. 6 displays the graphical depiction of these findings, illustrating how models react to data that has not been previously seen (testing data).

Table 6 shows the evaluation matrix on unseen samples (testing data) for all models using the under-sampled data.

Based on the results of the testing samples, the ensemble model with under-sampling and KNN were shown to be the least effective in predicting instances of fraud in the positive class. When comparing, the AdaBoost method achieved an accuracy of 0.970, a recall of 0.941, and F1-score values of 0.955. Fig. 7 displays the quantitative numbers to provide a more comprehensive comprehension of these outcomes.

Fig. 8 shows the ROC curve with the AUC-ROC values for all models. Significantly, AdaBoost and RF demonstrated the most excellent AUC-ROC values.

#### 4.3. Ensemble machine learning with SMOTE

Table 7 shows the evaluation results of the proposed Ensemble machine learning model (with SMOTE). In our test, the sampling strategy is set to 'auto' (default), which means SMOTE will resample the minority class until it has the same number of samples as the majority class (1:1 ratio).

Based on the results, RF demonstrated superior performance compared to the other models. It wrongly predicted 13 values, whereas the suggested ensemble model (with SMOTE) falsely predicted 39 values. Table 8 displays the accuracy values of all models.

The RF classifier achieved the most excellent accuracy (ACC) among all models in the testing sample, with a value of 0.99988. The ACC values of the Proposed ensemble model (with SMOTE) were closely followed, reaching 0.99964. Fig. 9 displays the graphical depiction of these findings, illustrating the accuracy of models to the testing data.

Table 9 shows the evaluation matrix applied to the testing dataset for all models to the balanced data (with SMOTE).

The results of the testing samples indicate that the suggested ensemble model (with SMOTE) and RF were the most effective models and achieved precision, recall, and an F1-score of 0.999, 1, and 0.999, respectively. By contrast, the AdaBoost method exhibits accuracy, recall, and F1-score values of 0.975, 0.947, and 0.961, respectively. Fig. 10 displays the quantitative data to provide a more comprehensive comprehension of these outcomes.

Fig. 11 exhibits the ROC curve with the respective AUC-ROC values for all models.

The suggested ensemble model and SMOTE, RF, and KNN demonstrated the most excellent value for the AUC-ROC).

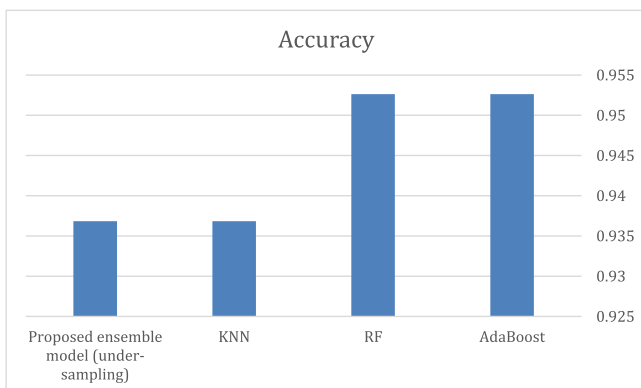


Fig. 6. Accuracy results comparative analysis of all models on the testing data.

Table 6

Evaluation matrix for the testing dataset (with under-sampling).

	Adaboost	RF	KNN	Proposed ensemble model (under-sampling)
precision	0.970	0.989	0.989	0.989
recall	0.941	0.922	0.892	0.892
f1-score	0.955	0.954	0.938	0.938

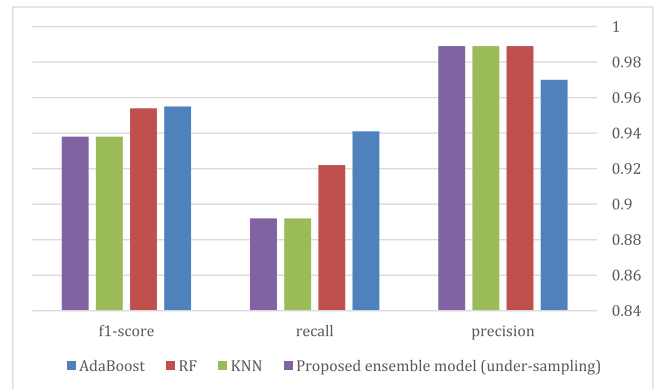


Fig. 7. Evaluation matrix values of the ensemble model (with under-sampling) for the testing dataset.

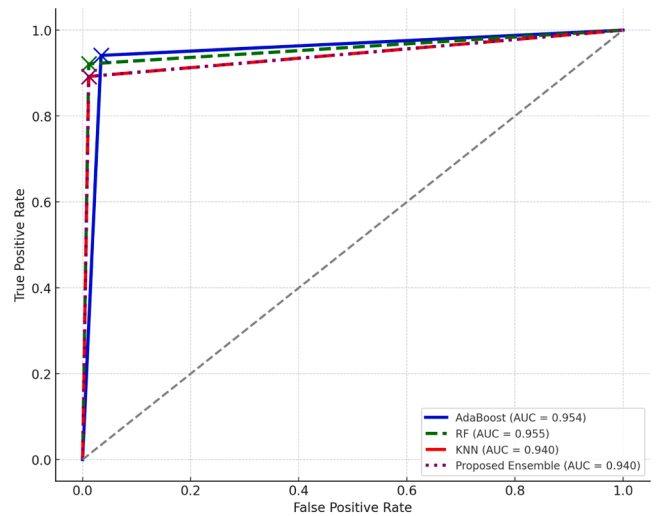


Fig. 8. ROC curve plots for the testing sample dataset.

Table 7

Confusion matrix values for the testing samples (with SMOTE).

Confusion Matrix	AdaBoost	RF	KNN	Proposed ensemble model (with SMOTE)
True Positives (TP)	52,094	55,003	55,003	55,003
True Negatives (TN)	53,729	55,067	54,968	55,034
False Positives (FP)	1344	13	105	39
False Negatives (FN)	2909	0	0	0

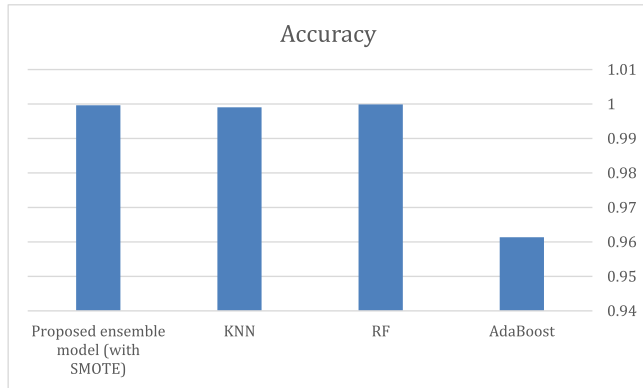
#### 4.4. Ensemble machine learning with SMOTE + ENN

Table 10 shows the confusion matrix findings from evaluating the proposed Ensemble machine learning model (using SMOTE + ENN). In

**Table 8**

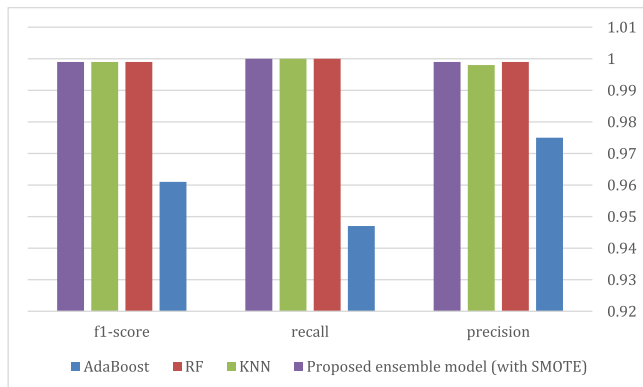
Test sample datasets accuracy results (with SMOTE).

	AdaBoost	RF	KNN	Proposed ensemble model (with SMOTE)
Accuracy	0.96136	0.99988	0.99904	0.99964

**Fig. 9.** Accuracy results comparative analysis of all models on the testing data.**Table 9**

Evaluation matrix for the testing dataset (with SMOTE).

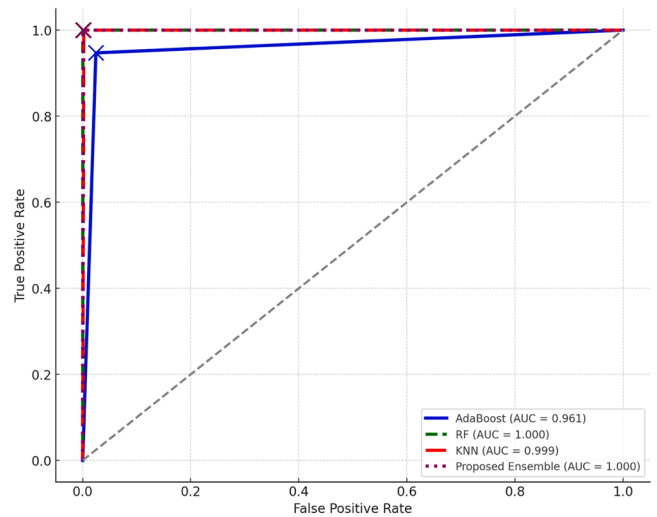
	AaBoost	RF	KNN	Proposed ensemble model (with SMOTE)
precision	0.975	0.999	0.998	0.999
recall	0.947	1	1	1
f1-score	0.961	0.999	0.999	0.999

**Fig. 10.** Evaluation matrix values of the ensemble model (with SMOTE) for the testing dataset.

our test, the sampling strategy is set to 0.5 to prevent excessive synthetic data generation before ENN is applied. Since ENN aggressively removes noisy and borderline samples, oversampling the minority class to 1.0 could lead to excessive data removal, reducing the dataset too much. By setting the sampling strategy to 0.5, the method maintains a balance between increasing minority class representation and preserving meaningful data after ENN refines the decision boundary, ensuring a cleaner and more effective resampling process.

Based on the results, the suggested ensemble model, which included the SMOTE + ENN techniques, made two incorrect predictions. Table 11 displays the precision measurements of each model on the testing samples.

The Proposed ensemble model (with SMOTE + ENN) classifier had the most excellent accuracy among all models in the testing data, scoring

**Fig. 11.** ROC curve plots for the testing sample dataset.**Table 10**

Confusion matrix values for the testing samples (with SMOTE + ENN).

Confusion Matrix	AdaBoost	RF	KNN	Proposed ensemble model (with SMOTE + ENN)
True Positives (TP)	25,551	27,416	27,420	27,420
True Negatives (TN)	54,320	55,050	55,033	55,050
False Positives (FP)	732	4	19	2
False Negatives (FN)	1869	4	0	0

**Table 11**

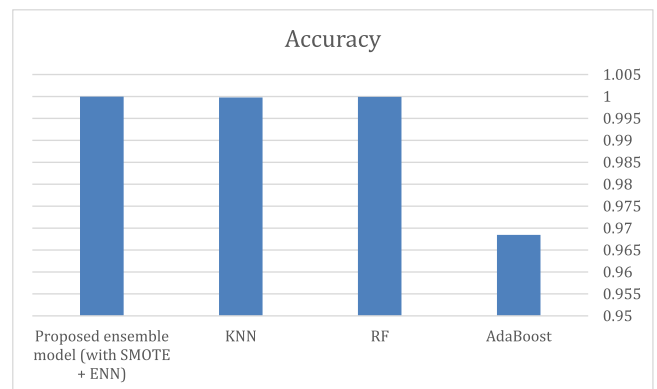
Test sample datasets accuracy results (with SMOTE + ENN).

	AdaBoost	RF	KNN	Proposed ensemble model (with SMOTE + ENN)
Accuracy	0.96846	0.99990	0.99976	0.99997

0.99997. The accuracy of RF classifiers closely followed, reaching a value of 0.99990. Fig. 12 illustrates the graphical depiction of these findings, showing how models react to the testing data.

Table 12 shows the evaluation matrix on unseen samples (testing data) for all models to the balanced data (with SMOTE + ENN).

The results of the testing samples show that the suggested ensemble

**Fig. 12.** Accuracy results comparative analysis of all models on the testing data.

**Table 12**

Precision, recall, and F1-score for the testing dataset (with SMOTE + ENN).

	AdaBoost	RF	KNN	Proposed ensemble model (with SMOTE + ENN)
precision	0.972	0.999	0.999	0.999
recall	0.932	0.999	1	1
f1-score	0.951	0.999	0.999	0.999

model, which uses SMOTE + ENN, and the KNN model are the most effective in predicting the fraud class. These models achieve accuracy, recall, and correspondingly an F1-score of 0.999, 1, and 0.999. By contrast, the RF display achieved high accuracy, recall, and F1-score, with values of 0.999. The results, as numbers are shown in Fig. 13, provide a more comprehensive comprehension of the results obtained.

Fig. 14 exhibits the ROC curve with the respective AUC-ROC values for all models. The proposed ensemble model, consisting of SMOTE + ENN, RF, and KNN, demonstrated the highest AUC-ROC value.

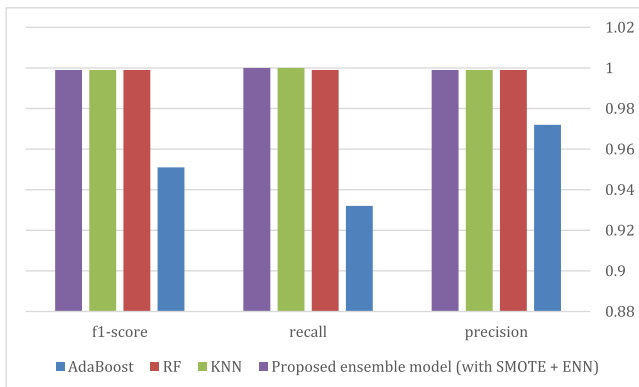
#### 4.5. Comparison with existing models

Several studies used SMOTE + ENN in credit card fraud detection to address data imbalance problems. These studies mainly concentrate on data preprocessing without extending their investigation to integrate SMOTE + ENN with classifiers, particularly ensemble methods.

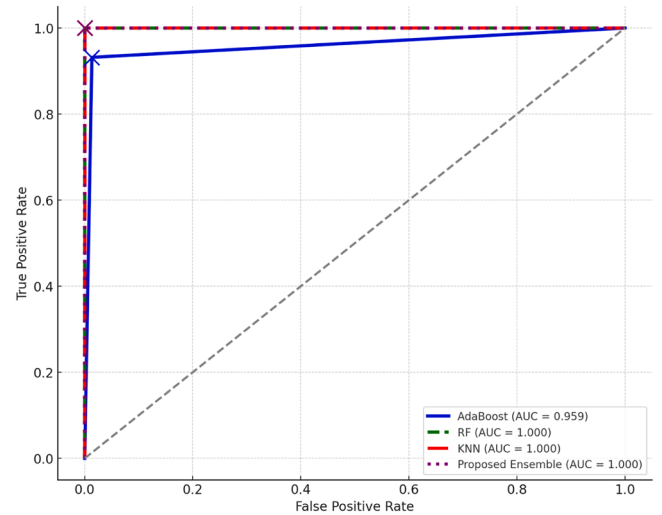
In addition, several previous studies have successfully combined SMOTE with ensemble classifiers for credit card fraud detection, demonstrating its effectiveness in improving performance by addressing data imbalance. Hence, no studies have explored the combination of SMOTEEN with ensemble classifiers. This paper addresses the gap by demonstrating the advantages of leveraging the SMOTEEN's hybrid data balancing approach in conjunction with ensemble classifiers.

The models presented in Sahithi et al. (2022); Khalid et al. (2024); Prusti and Rath, (2019) are similar to the proposed model. The model described in Prusti and Rath, (2019) included a KNN, Extreme Learning Machines (ELM), RF, Multilayer Perceptron (MLP), and Bagging classifier. The model suggested by Sahithi et al. (2022) included RF, KNN, (LR), Adaboost, and Bagging. In Khalid et al. (2024), SMOTE was used to address the class imbalance issue in the dataset, whereas SMOTE + ENN was used to address the class imbalance issue in our proposed model. Table 13 shows the results of the assessment measures used to compare the proposed model with previous studies.

The suggested model exhibited outstanding performance with an accuracy of 0.99997, surpassing the values obtained by references (Sahithi et al., 2022; Khalid et al., 2024; Prusti & Rath, 2019). This indicates that the suggested approach can appropriately categorize instances, such as improving proficiency in identifying fraudulent activities. The Proposed Model, which incorporates the ensemble



**Fig. 13.** Evaluation matrix values of the ensemble model (with SMOTE + ENN) for the testing dataset.



**Fig. 14.** ROC curve plots for the testing sample dataset.

**Table 13**

Evaluation of the proposed model when compared to previous studies.

	(Sahithi et al., 2022)	(Prusti & Rath, 2019)	(Khalid et al., 2024)	Proposed Model (with SMOTE + ENN)
Accuracy	0.99945	0.8383	0.99959	0.99997
precision	0.99947	0.945	0.9996	0.999
recall	0.99945	0.8647	0.9996	1
f1-score	0.99946	0.9031	0.9996	0.999

classifier with the SMOTE + ENN technique, shows exceptional performance, surpassing other models in accuracy, precision, and F1 score.

The Proposed Ensemble Model's ability to entirely eliminate false negatives underscores its power in reliably identifying all instances of fraud, thus enhancing the model's utility and trustworthiness in practical applications. The overall robustness of the proposed ensemble model, not only eliminates false negatives but also maintains a significantly lower false positive rate. This trade-off is crucial in fraud detection, as reducing false positives minimizes unnecessary transaction blocks while still ensuring all fraudulent cases are detected.

Through extensive experimentation, we show that this integrated approach enhances fraud detection accuracy by improving key metrics such as precision, recall, F1-score, and AUC-ROC. Our results confirm that training ensemble classifiers on a more balanced and cleaner dataset generated by SMOTE + ENN reduces false positives and false negatives, making fraud detection more reliable.

## 5. Conclusion and future work

The current study reveals significant obstacles in the field of CCFD. Selecting suitable classifiers is challenging due to the rapidly evolving landscape of machine learning algorithms and emerging fraud patterns, which require continuous adaptation and rigorous evaluation to maintain model effectiveness.

This paper evaluated the effectiveness of the presented models using a dataset from the real world. These efforts culminated in the model proposal combining AdaBoost, RF, and KNN inside a voting classifier and over-sampling using the SMOTE + ENN technique. The proposed model demonstrated strong performance and emphasized the effectiveness of merging numerous classifiers to improve fraud detection accuracy. Our experiments demonstrate that enhancing the ensemble model with SMOTE and ENN techniques leads to a significant improvement in the AUC scores for AdaBoost, Random Forest, and KNN classifiers. Our proposed model improves performance for challenges

such as fraudulent transaction detection regardless of the classification method used, as the AUC score generally provides a more accurate picture of the classification performance. Throughout the assessment process, the models were subjected to thorough testing, and their performance was carefully examined using a variety of measures.

The proposed model can be executed in real-life situations by integrating it into existing systems used by organizations such as banks and government agencies. The model would process live data streams, flag fraud, and send immediate alerts via reports or notifications to specified teams like IT staff. Reports with actionable insights would aid decision-making, and feedback processes with human supervision would refine the system over time.

In future work, we will evaluate the model's scalability by testing it on larger, more diverse real-world datasets and evaluate the potential deployment of the model in real-world financial infrastructures, analyzing its adaptability to live transactional data and its integration with existing fraud detection pipelines and ensuring its robustness across different financial environments. This will include assessing computational efficiency, scalability challenges, and the feasibility of real-time fraud detection in high-volume transactional systems. We also plan to integrate deep learning techniques to enhance fraud detection accuracy and adaptability to evolving patterns.

### Declaration of generative AI and AI-assisted technologies in scientific writing

During the preparation of this work, the authors used [ChatGPT] to [check the grammar and linguistic faults]. After using this tool, the authors reviewed and edited the content as needed. Therefore, they take full responsibility for the publication content.

### CRedit authorship contribution statement

**Khanda Hassan Ahmed:** Conceptualization, Methodology, Software, Data curation, Writing – original draft. **Stefan Axelsson:** Supervision, Validation. **Yuhong Li:** Writing – review & editing, Investigation, Validation. **Ali Makki Sagheer:** Writing – review & editing.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

This research received no specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.mlwa.2025.100675](https://doi.org/10.1016/j.mlwa.2025.100675).

### Data availability

The data is available online on Kaggle and listed in the reference list. The code will be available on request.

### References

- Annual Reports. (2024). April *UK Finance* <https://www.ukfinance.org.uk/annual-reports>.
- Azim Mim, M., Majadi, N., & Mazumder, P. (2024). A soft voting ensemble learning approach for credit card fraud detection. *Heliyon*, 10(3), Article e25466. <https://doi.org/10.1016/j.heliyon.2024.e25466>
- Bagga, S., Goyal, A., Gupta, N., & Goyal, A. (2020). Credit card fraud detection using pipelining and Ensemble learning. *Procedia Computer Science*, 173, 104–112. <https://doi.org/10.1016/j.procs.2020.06.014>
- Barongo, R. I., & Mbelwa, J. T. (2024). Using machine learning for detecting liquidity risk in banks. *Machine Learning with Applications*, 15, Article 100511. <https://doi.org/10.1016/j.mlwa.2023.100511>
- Bounab, R., Zarour, K., Guelib, B., & Khelifa, N. (2024). Enhancing medicare fraud detection through machine learning: Addressing class imbalance with SMOTE-ENN. *IEEE access : Practical innovations, open solutions*, 12, 54382–54396. <https://doi.org/10.1109/ACCESS.2024.3385781>. IEEE Access.
- Chhabra, R., Goswami, S., & Ranjan, R. K. (2024). A voting ensemble machine learning based credit card fraud detection using highly imbalance data. *Multimedia Tools and Applications*, 83(18), 54729–54753. <https://doi.org/10.1007/s11042-023-17766-9>
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), 6. <https://doi.org/10.1186/s12864-019-6413-7>
- Credit Card Fraud Detection. (2024). April <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>.
- Dileep, M. R., Navaneeth, A. V., & Abhishek, M. (2021). A novel approach for credit card fraud detection using decision tree and random forest algorithms. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (pp. 1025–1028). <https://doi.org/10.1109/ICICV50876.2021.9388431>
- Fourure, D., Javadi, M. U., Posocco, N., & Tihon, S. (2021). Anomaly detection: How to artificially increase your F1-score with a biased evaluation protocol. In Y. Dong, N. Kourtellis, B. Hammer, & J. A. Lozano (Eds.), *Machine learning and knowledge discovery in databases. applied data science track* (pp. 3–18). Springer International Publishing. [https://doi.org/10.1007/978-3-030-86514-6\\_1](https://doi.org/10.1007/978-3-030-86514-6_1)
- Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access : Practical Innovations, Open Solutions*, 10, 19083–19095. <https://doi.org/10.1109/ACCESS.2022.3151048>. IEEE Access.
- Meenakshi Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, 13(4), 1503–1511. <https://doi.org/10.1007/s41870-020-00430-y>.
- 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), 1. <https://doi.org/10.3390/bdcc8010006>. Article.
- Lokanan, M. E., & Sharma, K. (2022). Fraud prediction using machine learning: The case of investment advisors in Canada. *Machine Learning with Applications*, 8, Article 100269. <https://doi.org/10.1016/j.mlwa.2022.100269>
- Niveditha, G., Abarna, K., & Akshaya, G. V. (2019). Credit card fraud detection using random forest algorithm. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 5(2), 301–306. <https://doi.org/10.32628/CSEIT195261>
- Prasad, P. Y., Chowdary, A. S., Bavitha, C., Mounisha, E., & Reethika, C. (2023). A comparison study of fraud detection in usage of credit cards using machine learning. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1204–1209). <https://doi.org/10.1109/ICOEI56765.2023.10125838>
- Prusti, D., & Rath, S. K. (2019). Fraudulent transaction detection in credit card by applying ensemble machine learning techniques. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). <https://doi.org/10.1109/ICCCNT45670.2019.8944867>
- Qaddoura, R., & Biltawi, M. M. (2022). Improving fraud detection in an imbalanced class distribution using different oversampling techniques. In *2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEI)* (pp. 1–5). <https://doi.org/10.1109/EICEEI56378.2022.10050500>
- Raghavan, P., & Gayar, N. E. (2019). Fraud detection using Machine Learning and Deep Learning. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)* (pp. 334–339). <https://doi.org/10.1109/ICCIKE47802.2019.9004231>
- Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. *Procedia Computer Science*, 148, 45–54. <https://doi.org/10.1016/j.procs.2019.01.007>
- Sahithi, G. L., Roshmi, V., Sameera, Y. V., & Pradeepini, G. (2022). Credit card fraud detection using ensemble methods in Machine learning. In *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1237–1241). <https://doi.org/10.1109/ICOEI53556.2022.9776955>
- Sailusha, R., Gnanaswar, V., Ramesh, R., & Rao, G. R. (2020). Credit card fraud detection using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1264–1270). <https://doi.org/10.1109/ICICCS48265.2020.9121114>
- Saputra, A. (2019). Fraud detection using machine learning in e-commerce. *International Journal of Advanced Computer Science and Applications*, 10(9). <https://doi.org/10.14569/IJACSA.2019.0100943>
- Shah, V., & Passi, K. (2021). Data balancing for credit card fraud detection using complementary neural networks and SMOTE algorithm. In N. Chaubey, S. Parikh, & K. Amin (Eds.), *Computing science, communication and security* (pp. 3–16). Springer International Publishing. [https://doi.org/10.1007/978-3-030-76776-1\\_1](https://doi.org/10.1007/978-3-030-76776-1_1)
- Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access: Practical Innovations, Open Solutions*, 8, 25579–25587. <https://doi.org/10.1109/ACCESS.2020.2971354>. IEEE Access.
- Tanouz, D., Subramanian, R. R., Eswar, D., Reddy, G. V. P., Kumar, A. R., & Praneth, C. V. N. M. (2021). Credit card fraud detection using machine learning. In

- 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 967–972). <https://doi.org/10.1109/ICICCS51141.2021.9432308>
- Tekkali, C. G., & Natarajan, K. (2024). Assessing CNN's performance with multiple optimization functions for credit card fraud detection. *Procedia Computer Science*, 235, 2035–2042. <https://doi.org/10.1016/j.procs.2024.04.193>
- Vairam, T., Sarathambekai, S., Bhavadharani, S., Kavi Dharshini, A., Nithya Sri, N., & Sen, T. (2022). Evaluation of naïve bayes and voting classifier algorithm for credit card fraud detection. In , 1. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 602–608). <https://doi.org/10.1109/ICACCS54159.2022.9784968>
- Wahab, F., Khan, I., & Sabada, S. (2024). Credit card default prediction using ML and DL techniques. *Internet of Things and Cyber-Physical Systems*, 4, 293–306. <https://doi.org/10.1016/j.iotcps.2024.09.001>
- Yacoub, R., & Axman, D. (2020). Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In S. Eger, Y. Gao, M. Peyrard, W. Zhao, & E. Hovy (Eds.), *Proceedings of the first workshop on evaluation and comparison of nlp systems* (pp. 79–91). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.eval4nlp-1.9>.