mathematics

MDPI

Article

Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture

Esraa Faisal Malik [1], Khai Wah Khaw [1], Bahari Belaton [2], Wai Peng Wong [3] and XinYing Chew [2,*]

[1] School of Management, Universiti Sains Malaysia, Gelugor 11800, Malaysia; esraa.f@student.usm.my (E.F.M.); khaiwah@usm.my (K.W.K.)
[2] School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Malaysia; bahari@usm.my
[3] School of Information Technology, Monash University, Malaysia Campus, Subang Jaya 47500, Malaysia; waipeng.wong@monash.edu
* Correspondence: xinying@usm.my

Abstract: The negative effect of financial crimes on financial institutions has grown dramatically over the years. To detect crimes such as credit card fraud, several single and hybrid machine learning approaches have been used. However, these approaches have significant limitations as no further investigation on different hybrid algorithms for a given dataset were studied. This research proposes and investigates seven hybrid machine learning models to detect fraudulent activities with a real word dataset. The developed hybrid models consisted of two phases, state-of-the-art machine learning algorithms were used first to detect credit card fraud, then, hybrid methods were constructed based on the best single algorithm from the first phase. Our findings indicated that the hybrid model Adaboost + LGBM is the champion model as it displayed the highest performance. Future studies should focus on studying different types of hybridization and algorithms in the credit card domain.

Keywords: classification; credit card; data mining; fraud detection; hybrid; machine learning

MSC: 68T01

Citation: Malik, E.F.; Khaw, K.W.; Belaton, B.; Wong, W.P.; Chew, X. Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture. *Mathematics* **2022**, *10*, 1480. https://doi.org/10.3390/math10091480

Academic Editors: Andrea Prati, Luis Javier García Villalba and Vincent A. Cicirello

Received: 25 January 2022
Accepted: 27 April 2022
Published: 28 April 2022

1. Introduction

The rapidly changing world and the evolving financial industry have led to ease in an individual's life, especially in the time of COVID19 during which many have shifted to online platforms. Consequently, financial crimes such as credit card fraud have significantly risen. Since 2011, there has been a rapid growth in global losses due to payment fraud as it jumped from USD 9.84 billion in 2011 to USD 32.39 billion in 2020 and it will inevitably be a serious worldwide predicament as it is expected to cost USD 40.62 billion in 2027 [1]. This problem has captured the concern of governments and financial institutions, not only because of the monetary losses but because these acts can seriously harm a nation's reputation. Black's Law Dictionary defines fraud as "A knowing misrepresentation of the truth or concealment of a material fact to induce another to act to his or her detriment" [2]. There are various types of fraud including but not limited to, tax evasion, insurance fraud, credit card fraud, money laundering and identity theft. Most banks and financial firms use rule-based systems, in which an expert will use historical fraud data to define a set of rules, and a system will raise an alarm if a new transaction matches one of the rules [3,4]. The main limitations of this manual process are that it is reactive, lacks flexibility and consistency as well as the fact that it is time-consuming [5]. Amid these challenges, firms ought to espouse a proactive technology-driven approach for fraud detection, particularly with the new sophisticated criminal techniques that are continually evolving with technological advancements. The era of technology advancement has aided the financial industry in the better detection of these financial crimes by harnessing the power of machine learning techniques that can uncover hidden patterns and, therefore, identify fraudulent financial

activities using the realistic dataset to simplify decision-making processes. Furthermore, it helps in keeping up with the ever-changing sophisticated fraud techniques.

Machine learning is the science of getting computers to learn without being explicitly programmed [6]. It has been commonly used in a wide range of disciplines such as; Chemistry [7], Bioinformatics [8], Manufacturing industries [9], the Medical Field [10–12], Biology [13] and in Finance [3,14–16].

For fraud detection, machine learning is mainly employed to help organizations and financial institutions better detect fraudulent transactions. However, fraud detection can pose a challenge for machine learning for several reasons [14]:

- The distribution of the data is highly imbalanced as the number of fraudulent transactions is very small.
- The data is continually evolving over time.
- Lack of real-world dataset due to privacy concerns.

As an attempt to overcome these challenges, multiple approaches were proposed in the literature while the main focus was placed on utilizing the idea of a hybrid model.

Several studies on the formation of hybrid models for fraud detection have been reported [13,16–21]. However, these hybrid models only utilized a single model without the consideration of the performance of other models to confirm that the selected model is the optimum choice for the chosen dataset. As such, this has inadvertently led to results inaccuracy and a lack of generalization for the model. The key contribution of this paper is to develop and investigate the use of multiple hybrid models for the same dataset and determine a Champion Hybrid Model based on the evaluation of their performance prediction. We examine the combination of the following eight supervised machine learning algorithms: linear regression (LR), support vector machine (SVM), Naïve Bayes (NB), random forest (RF) decision tree (DT), Light Gradient Boosting Machine (LGBM) and eXtreme Gradient Boosting (XBGOOST) and Adaptive Boosting (Adaboost). The scope of the study is limited to classification supervised machine learning in credit card fraud detection as the nature of most fraud datasets, specifically credit card datasets, is labelled. A real-world dataset was used, and several evaluation metrics were adopted to assess and compare the prediction performance of the proposed hybrid models and the state-of-the-art machine learning algorithms.

The remainder of this paper is organized as follows. The second section provides a brief overview of the hybrid models in the fraud detection domain the literature. The third section discusses in detail the methods and materials including the data collection and preparation. In Section 4, we discuss the model development using our proposed models and model evaluation. The results and discussion are outlined in Section 5. Finally, the conclusion and future works are presented in Section 6.

2. Related Works

Fraud detection has received much attention in the past decade. In this section, hybrid machine learning algorithms used in the credit card fraud domain are reviewed. A growing body of literature has proposed approaches with which to enhance fraud detection.

Combing different approaches together, the authors in [3] investigated a combination of different approaches together as they proposed a new voting mechanism called OPWEM, standing for; optimistic, pessimistic, and weighted voting in an ensemble of models that can work in tandem with rule-based systems. The authors' use of OPWEM is fully justified as they suggest that depending on the bank's strategy for false alarm rates, the bank management should choose one of the voting techniques. For example, pessimistic voting (PES) should be chosen if a bank desires to locate as many fraud cases as possible. On the other hand, a bank that strives for a low false alarm rate should use the optimistic voting (OPT) strategy. Additionally, weighted voting (WGT) discovered more frauds than OPT with a marginal false alarm rate. Therefore, it may be selected as a good alternative to OPT and PES. Additionally, a hybrid framework model was presented based on the combination of unsupervised and supervised learning models by the author in [17]. The author's

objective was to identify fraudulent transactions at a low cost including the amount of time and effort spent by bank practitioners to reach the necessary level of expertise in machine learning classification methods. The author employed a straightforward approach for one-class classification, with the improvement that the data description boundary is altered based on the account holders' purchasing behavior. To enhance the model's output, a post-processing operation was implemented in which rule-based filters are used to pass the flagged accounts. The author concluded that the one-class classification method is highly suitable for complex and large-scale datasets of transaction data as it assists in developing an account group structure that provides personalized models for different types of cardholder behavior. It has been mentioned by the author that the used technique, combined with the post-processing level of the rule-based filters, yields the best results. However, the main limitation of this study is that the experimental findings display that most of the fraudulent cases detected using the hybrid technique are missed by the bank's rule-based system, and vice versa. This implies that both methods should be used concurrently to gain the optimum results.

Moreover, a hybrid model for improving fraud detection accuracy by combining supervised and unsupervised methods was presented by the authors in [22]. They displayed several criteria for calculating outlier scores at various levels of granularity (from high granular card-specific outlier scores to low granular global outlier scores). Then, they evaluated their added value in terms of precision once integrated as characteristics in a supervised learning approach. Unfortunately, in terms of local and global methods, the results are unconvincing. However, the model provides a more considerable result in terms of Area Under the Precision–Recall Curve (AUC-PR).

The authors in [23] applied Bayesian Classification and Association Rule Learning (ARL) to investigate and discover the real transaction signs of fraudulent accounts, and to provide a reference in fraud prevention to the financial industry. Based on these signs, a fraudulent account detection system was developed, and the signs were further investigated by utilizing real-time daily transaction data. They concluded that the proposed method of their study is effective and efficient and can be used by financial institutions to minimize the need for the manual screening of fraudulent accounts. Likewise, an intelligent model for credit card fraud detection to identify fraud in anonymous and heavily skewed credit card datasets was proposed by the authors in [20]. The authors divided each customer's transactions into fraudulent and legitimate transactions, then they applied the Apriori algorithm to both sets to determine the patterns for fraudulent and legitimate transactions. Consequently, to detect fraud, they suggested a matching algorithm that searches pattern databases for a match with the incoming transaction. Another important point to note is that to deal with the data's anonymity each feature was treated equally when looking for patterns and therefore no preference was given to any feature. Finally, the authors suggested running the proposed model at fixed time points occasionally to upgrade the legal and fraud pattern database as a result of customer fraudulent behavior changing slightly over time.

Similarly, the authors in [21] also presented a hybrid model that combines ARL and process-mining by conducting a process-mining inquiry to collect a number of fraud variables to create some association rules for fraud detection. The aim of process-mining in this context is to inspect skipped tasks, resources, throughput time, and decision points based on simple rules in the Standard operating procedure (SOP). In the first phase, they used a process-mining technique to extract the variables of fraudulent cases from the dataset. Then, an expert determines whether a case contains fraud variables. In the second phase, an Apriori algorithm is used to produce either fraud cases or legal cases. Eventually, as the detection rules, only the association rules with specific consequences such as expert judgement regarding fraudulent status are selected.

Furthermore, a twelve-machine learning algorithm in conjunction with the AdaBoost and majority voting methods using a real credit card dataset obtained from a financial institution has been used to investigate the performance of the used classifiers [16]. Their

result for the highest Matthews correlation coefficient (MCC) score was 0.823, which was obtained by a majority of the votes. However, when using AdaBoost and majority voting procedures, a perfect MCC score of 1 was obtained. To further assess the hybrid models, noise ranging from 10 percent to 30 percent was added to the data samples. When 30 percent noise was added to the data set, the majority voting procedure produced the best MCC score of 0.942. Therefore, the authors reported that the majority vote method performs well in the presence of noise.

More recent research was conducted to develop a hybrid model to detect credit card fraud using credit card datasets and utilizing machine learning classifiers with LR, Gradient Boosting (GB), RF and voting classier [24]. The author found that RF and GB gave maximum detection rates of 99.99 percent. Although all the aforementioned studies were concerned with fraud detection, different algorithms were used depending on the nature of the dataset. As evident from previous efforts, various approaches were used to detect fraudulent transactions in the financial sector especially the credit card domain either using a single machine learning algorithm or hybrid models. However, these hybrid models only utilized a single model without consideration of the performance of other models to confirm that the selected model is the optimum choice for the chosen dataset. Therefore, this might inadvertently lead to inaccurate results and a lack of generalization for the proposed model. Therefore, a comparison of several hybrid models using the same datasets is still needed to understand the relative performance of the proposed technique. The key contribution of this paper is to develop and investigate the use of multiple hybrid models for the same dataset and determine a champion hybrid model based on the evaluation of their performance prediction.

3. Materials and Methods

3.1. Data Collection

The dataset in this research was provided by Vesta Corporation and it was released in the Kaggle community by researchers from IEEE Computational Intelligence Society (IEEE-CIS) [25]. This dataset contains about half a million credit card transactions with a target feature and 432 features for each transaction where it varies between numeric and categorical features. The data are highly imbalanced at around 569 K of legitimate transactions and 20 K of fraudulent transactions, hence the imbalance rate is around 0.035.

Initially, the original dataset included many transactions, thus, to avoid computational cost and model training delay, a smaller random subset sample was created from the original dataset to prove our concept of the study (POC). However, the POC dataset still suffers from the same imbalance ratio as the original dataset. Figure 1 illustrates a sample of the dataset. Due to the large number of features, not all features are included in the figure, only the first and last are included for some features with the same first name. For instance, card features include around six features, card 1, card 2, card 3 card 4 and card 6.

TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card6	addr1	addr2	dist1	dist2	P_emaildomain	R_emaildomain	C1	C14	D1	D15	M1	M9	V1	V339	id_01	id_38	DeviceType	DeviceInfo
2,987,000	0	86,400	68.5	W	13,926	credit	315	87	19				1	1	14	0	T		1		0	T	mobile	SAMSUNG SM/G892A Build/NRD90M
2,987,002	0	86,469	59	W	4663	debit	330	87	287		outlook.com		1	1	0	315	T	F	1		-5	T	desktop	Windows
2,987,003	0	86,499	50	W	18,132	debit	476	87			yahoo.com		2	1	112	111					-5	T	desktop	
2,987,004	0	86,506	50	H	4497	credit	420	87			gmail.com		1	1	0					0	0	T	desktop	MacOS
2,987,005	0	86,510	49	W	5937	debit	272	87	36		gmail.com		1	1	0	0	T		1		-5	T	desktop	Windows
2,987,006	0	86,522	159	W	12,308	debit	126	87	0		yahoo.com		1	1	0	0	T	T	1		-15			
2,987,007	0	86,529	422.5	W	12,695	debit	325	87			mail.com		1	1	0	0					0	T	mobile	
2,987,008	0	86,535	15	H	2803	debit	337	87			anonymous.com		1	1	0					0	-10	T	desktop	Windows
2,987,009	0	86,536	117	W	17,399	debit	204	87	19		yahoo.com		2	2	61	318	T		1		-5	T	desktop	Windows

Figure 1. Sample of the used dataset before prepossessing phase.

3.2. Data Preparation

In any predictive analytics study, this phase is the most critical one as it defines the success of the study. Real-world datasets are well-known to be chaotic because they contain a massive number of outliers, missing values, irregular cardinality, etc. Such phenomena can lead to the failure of the research if not handled correctly. In this paper, our preparation includes handling missing values, transforming categorical features, feature scaling, feature selection and resampling.

3.2.1. Missing Values

As can be seen in Figure 2, the missing values in some attributes reached up to 99 percent. The whole dataset shows around 45 percent of the missing values, and imputing such large amounts causes a bias in the model and induces incorrect predictions. Therefore, as a rule of thumb, if the missing values reaches above 60 percent, the features will be removed as the amount of information stored in that specific feature is insufficient and will have no contribution to the prediction model [26]. The rest of the features, which represent less than 50 percent of the missing values, were imputed using the mode for categorical and the median for numeric features. This is due to the data are highly imbalanced, thus the median is a better option than the mean.

	Missing Ratio
id_24	99.196159
id_25	99.130965
id_07	99.127070
id_08	99.127070
id_21	99.126393
...	...
V241	77.913435
V229	77.913435
V217	77.913435
V223	77.913435
V224	77.913435

Figure 2. Missing ratio for each column in the dataset.

3.2.2. Encoding Categorical Features

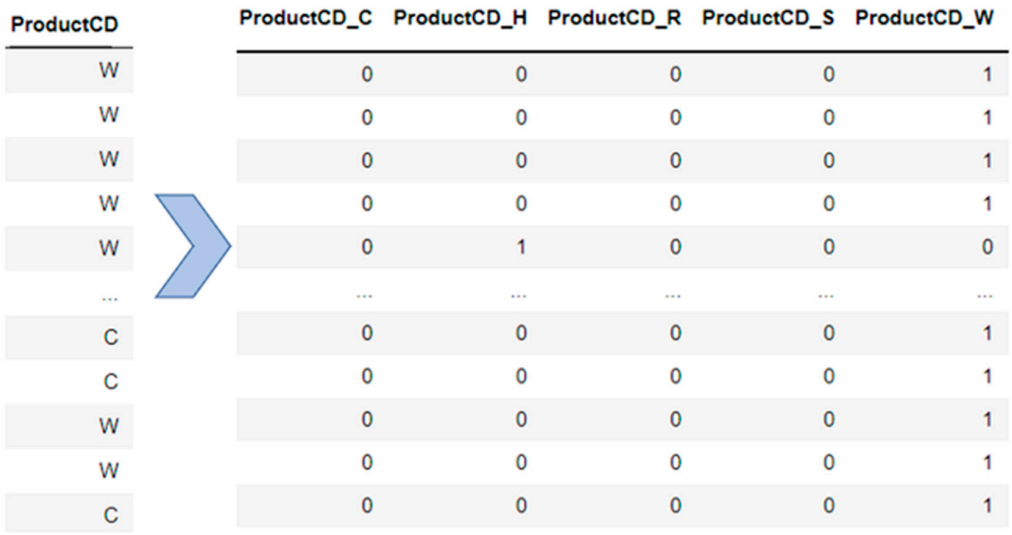
Most machine learning algorithms require the input and output features to be in a numerical format. This implies that categorical data must be converted to numbers before the development of a prediction model. After removing the features with high missing values, 15 categorical features remained. Ten of them contained only two levels of cardinality (e.g., attributes with false and true); therefore, they were replaced by 0 and 1, respectively. The remainder of the categorical features have more than two levels, and were instead converted using the one-hot encoding technique which is a straightforward and widely used encoding method [27]. Each category value is converted into a new categorical column and given a binary value of 1 or 0. An illustration of how one-hot encoding works is provided in Figure 3. In this example, the categorical attribute Product CD with the categories C, H, R, S, and other as W is encoded with 5-dimensional feature vectors [1, 0, 0, 0, 0], [0, 1, 0, 0, 0], [0, 0, 1, 0, 0], [0, 0, 0, 1, 0] and [0, 0, 0, 0, 1]. This technique is implemented using get dummies in Pandas software library.

3.2.3. Feature Scaling

Most features of a real-world dataset will vary in range, unit and magnitude. A problem arises when one feature's magnitude is higher than the rest, as it will then naturally dominate other features. As a consequence, raw data should be scaled to fit classification algorithms and eliminate the impact of various quantitative units [28]. Therefore, in this research, the MinMaxScaler technique was used to rescale the features between 0 and 1. The benefit of this technique is that it is robust to outliers as it uses statistics techniques that do not affect the variance of the data (Equation (1)) [9].

$$x' = (x - \min(x))/\max(x) - \min(x) \quad (1)$$

As shown from the above Equation (1), x represents the original value, x' represents the scaled value, max indicates the upper bound of the feature value, and min represents the lower bound. For data with numerous zero entries, MinMaxScaler scaling preserves the sparsity of the input data and thus saves time as a result [29].



ProductCD	ProductCD_C	ProductCD_H	ProductCD_R	ProductCD_S	ProductCD_W
W	0	0	0	0	1
W	0	0	0	0	1
W	0	0	0	0	1
W	0	0	0	0	1
W	0	1	0	0	0
...	...	...	...	...	...
C	0	0	0	0	1
C	0	0	0	0	1
W	0	0	0	0	1
W	0	0	0	0	1
C	0	0	0	0	1

Figure 3. One-hot encoding.

3.2.4. Feature Selection

Feature selection is an essential task to perform in data preparation as the curse of dimensionality reduction is a serious issue that might lead to overfitting. Feature selection works by eliminating redundant and irrelevant features. The most famous techniques are filter and wrapper, with each technique having its own merits and limitations. The wrapper approach has a few limitations such as its dependence on the algorithm as an evaluation function to choose the features in addition to its high computational cost [30]. On the other hand, the limitation of the filter approach is that it looks for features as individuals, therefore, features with high dependency on one other will be missed by this approach when combined with other features. An alternative solution, which is less exhaustive and has fewer shortcomings, is to use a hybrid feature selection approach which integrates both filter and wrapper approaches [31]. More specifically, a correlation-based filter was used to test the correlation between the numerical features. As shown in Figure 4, some group features show high correlation results. Accordingly, strongly positive features (0.8) that correlated with each other were removed because they would make no contribution to the prediction model, prevent overfitting, and doing so would save computation resources [32].

Following this, a wrapper method called SVM-Recursive Feature Elimination (SVM-RFE) was applied as it is one of the most common techniques for feature selection [33]. It was tested on a validation dataset using a loop with a range of 10 to 140 features; as a result, the initial number of features with the highest performance was chosen, which is equal to 80 features.

3.2.5. Data Resampling

The dataset in this paper is highly imbalanced; the problem with an imbalanced dataset is that most of the machine learning algorithms impose the assumption of equal distribution for both minority and majority classes which provides mispresented results and poor predictive modelling performance. Furthermore, the imbalance problem appears to be related to learning with too few minorities class examples in the presence of other complicating factors, such as class overlapping [34–36]. One popular way to deal with such a condition directly is the Synthetic Minority Over-Sampling Technique with Edited Nearest Neighbors (SMOTE-ENN) which works by firstly applying SMOTE for the oversampling phase followed by ENN as a data-cleaning method to eliminate the overlapping between

classes to produce better-defined class clusters. Due to this benefit, SMOTE-ENN has been applied in this research to overcome the problem of imbalanced datasets and overlapping classes. Table 1 illustrates the number of observations for each class (fraud and non-fraud) before and after SMOTE-ENN. As a result, 37, 894 observations for fraudulent as well as non-fraudulent cases are taken in this study. Furthermore, Figure 5 displays a sample of the dataset after the preprocessing phase.

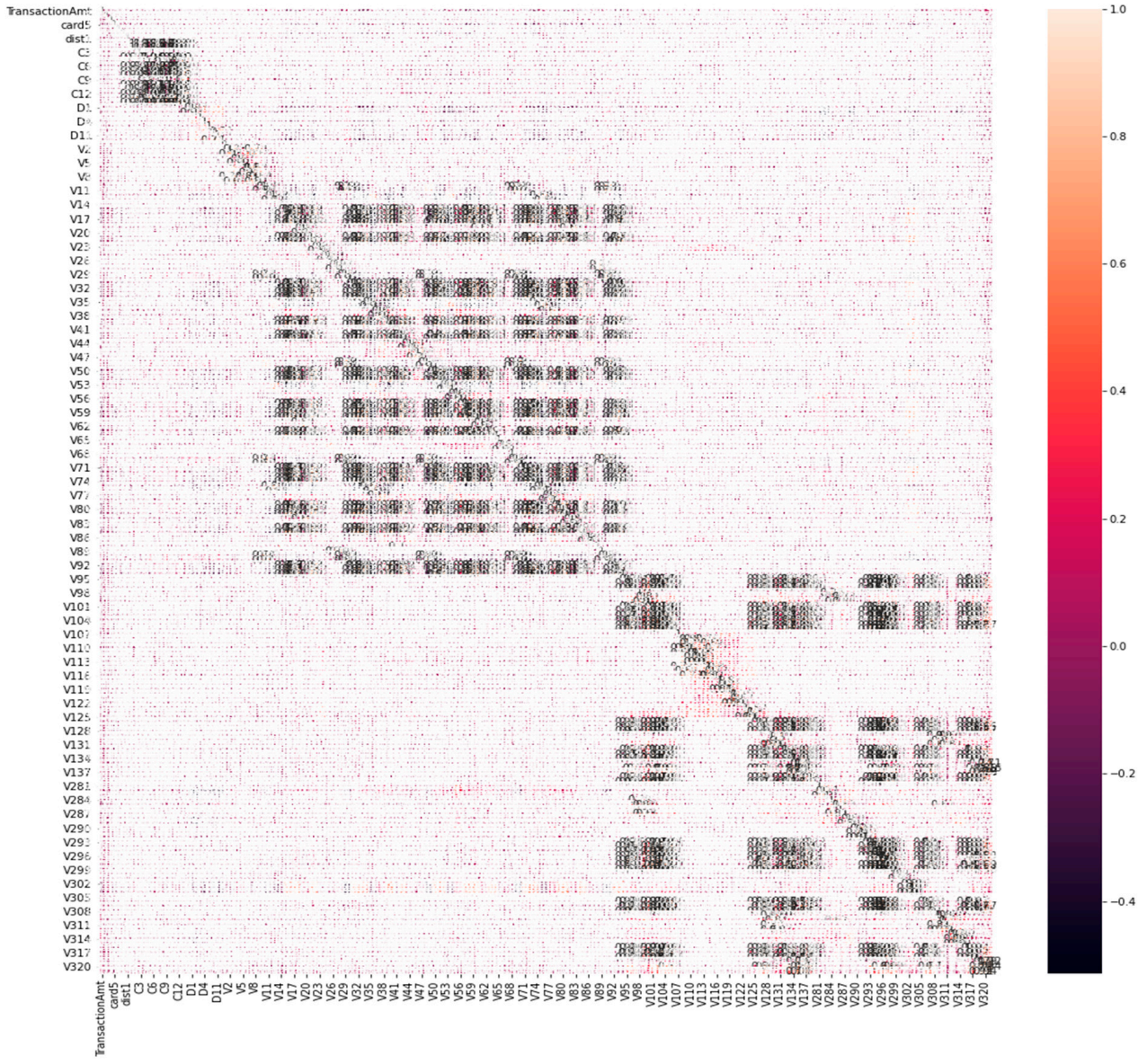


Figure 4. Correlation-based filter.

TransactionAmt	C3	D1	D3	D4	D5	D10	D11	D15	V4	V6	V8	V10	V14	V15	V19	V23	V25	V27	V29	V35	V37	V39	V44
0.027980932	0	0	0.0102	0.12373	0.01368	0.52968	0	0.53129	0.16667	0.125	0.16667	0	1	0	0.14286	0.08333	0.25	0	0.25	0.33333	0.02703	0	0.02857
0.004638009	0	0.22188	0.18112	0.12373	0.01368	0	0.21386	0.16496	0.16667	0.125	0.16667	0	1	0	0.14286	0.08333	0.25	0	0	0.33333	0.02703	0	0.02857
0.0014877	0	0.10781	0.00893	0.19371	0.00958	0	0.06476	0.08191	0.16667	0.125	0.16667	0	1	0.14286	0.14286	0.08333	0.25	0	0	0	0.02703	0.07143	0.02857
0.017614702	0	0	0.0102	0.12677	0.0041	0.00342	0.00452	0.00683	0.16667	0.125	0.16667	0.33333	1	0	0.14286	0.08333	0.25	0	0.25	0.33333	0.02703	0	0.02857
0.032748149	0	0	0.0102	0.1501	0.01368	0.01712	0.06476	0.06257	0.16667	0.125	0.16667	0	1	0	0.14286	0.08333	0.25	0	0	0.33333	0.02703	0	0.02857
0.02452881	0	0	0.0102	0.1501	0.01368	0.01712	0.06476	0.06257	0.16667	0.125	0.16667	0	1	0	0.14286	0.08333	0.25	0	0	0.33333	0.02703	0	0.02857
0.019104046	0	0.50938	0	0.45538	0	0.37329	0.41867	0.37543	0.16667	0.125	0.16667	0	1	0	0.14286	0.08333	0.25	0	0	0	0.02703	0	0.02857
0.030446734	0	0	0	0.12373	0	0	0	0.00341	0.16667	0.125	0.16667	0	1	0	0.14286	0.08333	0.25	0	0	0.33333	0.02703	0	0.02857
0.004638009	0	0	0.0102	0.12373	0.01368	0	0	0.00341	0.16667	0.25	0.33333	0.33333	1	0	0.14286	0.08333	0.25	0	0	0.33333	0.02703	0	0.02857

Figure 5. Sample of the used dataset before prepossessing phase.

Table 1. The number of observations for each class (fraud and non-fraud) before and after SMOTE-ENN.

	Number of Observations for Fraud Cases	Number of Observations for Non-Fraud Cases
Before SMOTE-ENN	1157	32,060
After SMOTE-ENN	28,689	27,155

4. Model Development

Different machine learning classification techniques have been applied to detect fraudulent transactions as discussed earlier. Yet, there is no optimal algorithm for a specific problem [26]. Therefore, eight different linear and nonlinear algorithms were selected from the literature as they indicated promising performance in the context of fraud detection [37], including LR, RF, DT, XGBOOST, SVM, NB, Adaboost and LGBM.

In our study, we undertook the methodology suggested by [38] that has been applied in credit rating, in the credit card fraud detection domain for the first time. Additionally, the majority of the algorithms used in this research were different from the research in [39]. The development phase of the hybrid models is divided into two phases. In the first phase, a single baseline machine learning classification model was developed using the following eight machine learning algorithms: LR, RF, DT, XGBOOST, SVM, NB, Adaboost and LGBM where their performance was investigated.

Even though algorithm parameter tuning can be useful, a consideration of default parameters is more common in practice. The need for considerable work and time for tuning can dissuade people from implementing the step and could also lead to issues of overfitting for specific datasets [40]. Appropriately, there were no deliberate efforts to fine-tune the parameters of the methods.

Subsequently, in the second phase of the proposed model, the algorithm with the best performance from the previous experiment based on the highest Area Under the Receiver Operating Characteristic (AUROC) metric served as a baseline model and was used to train the rest of the seven algorithms. The correctly classified data points—true positive (TP) and true negative (TN)—that are generated by the single machine learning algorithm with the highest performance in level one were used to train the hybrid models separately. Consequently, seven hybrid models were constructed and are as follows:

(1) The best baseline single model + LR;
(2) The best baseline single model + RF;
(3) The best baseline single model + DT;
(4) The best baseline single model + XGBOOST;
(5) The best baseline single model + SVM;
(6) The best baseline single model + NB;
(7) The best baseline single model + LGBM.

The utilization of the algorithms, which is derived by its score in AUROCmetric for detecting the correct classes, will assist the hybrid models to precisely detect fraudulent and legitimate activities. The proposed hybrid models will be compared with state-of-the-art algorithms to check their effectiveness. Figure 6 presents the details of the proposed flowchart.

According to the No Free Lunch Theorem, no single model or algorithm can handle all classification problems [26,28]. Furthermore, each different algorithm has its advantages and disadvantages as illustrated in Table 2 [16,41–44]. Consequently, the combination of several algorithms exploits the weaknesses of a single one, such as overfitting. This combination of several algorithms will be beneficial if the algorithms are substantially different from each other. Combining these algorithms together will result in optimal performance and help to overcome the limitation of a single classifier and therefore enhance the detection of fraudulent cases. This distinction could be in the algorithm, or the data used in that algorithm.

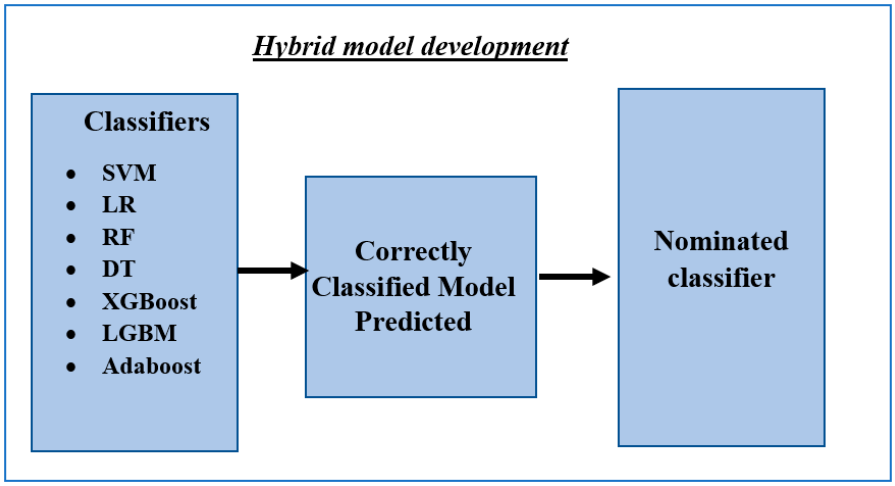


Figure 6. Flowchart of the proposed model.

Table 2. Comparison between different algorithms' strengths and limitations.

Algorithm	Strength	Limitation
LR	Simple parametric approach and ease of implementation	Poor classification performance
DT	Easy to understand and interpret, not requiring complicated data preparation, provides a strong indicator of which features are most relevant for the classification model	Vulnerable to overfitting and impossibility of performing prediction for numeric labels
NB	Robust to noisy and irrelevant data points, computationally efficient and easily understood	The Independence assumption may not hold for some features
SVM	High tolerance to noisy features, effective in high dimensional spaces and memory-efficient	Computationally expensive in training and weak interpretability
LGBM	High training speed and performance and low memory utilization	Has a high chance of overfitting
Adaboost	Ease of use, less parameter tweaking and less susceptible to overfitting	Sensitive to outliers and noisy

5. Model Evaluation

Stratified k-folds validation was applied to measure the efficiency of the proposed model in which it attempts to ensure that both classes (fraud and non-fraud) are roughly evenly distributed in each fold [27]. In this research, we employed five k, where the validation set is randomly divided into five equal-sized subsets. At each phase of validation, a subset of 25 percent was set aside as the validation dataset to assess the output of the proposed method, while the remaining four subsets that encompass 75 percent were used as a training set.

We employed various performance evaluation metrics that have been widely seen in the literature. It should be noted that the accuracy score is inadequate in the case of a highly imbalanced dataset owing to the overwhelming majority class. Consequently, different criteria are needed to evaluate the model's performance such as AUROC, AUC-PR, Type-I error, Type-II error F1-measure, recall, precision, misclassification rate and Specificity or True Negative Rate (TNR). The terms used in the applied metrics are defined as follows [28]:

- True positive (TP) implies the number of correctly classified data as fraudulent credit card transactions.
- True negative (TN) implies the number of correctly classified data as legitimate credit card transactions.
- False positive (FP) denotes the number of legitimate credit card transactions classified as fraudulent.

- False Negative (FN) denotes the number of fraudulent credit card transactions classified as legitimate.

Although there is no ultimate individual evaluation metric that can be used to evaluate both negative and positive classes, it was decided that the best overall performance metric for the imbalanced fraud dataset was to use AUROC [29]. AUROC is an evaluation classification metric that is used to calculate the area under the ROC curve, which gives equal consideration to positive and negative classes. The ROC curve presents a compromise between the true positive rate (TPR) and false-positive rate (FPR) and it is calculated as follows:

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \tag{2}$$

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) \tag{3}$$

The AUROC values vary from 0 to 1 where 1 represents ideal prediction, 0 represents terrible prediction performance and 0.5 represents random performance. The advantage of AUROC is that it does not require a specific cut off value. Additionally, it provides valuable information on whether the model is indeed obtaining knowledge from the data or simply guessing. Additionally, it can be more readily understood compared with the numerical methods due to its visual representation method.

In addition, recall and precision were also suitable to evaluate the predictive model to check if it is capable to identify fraudulent transactions accurately. A recall which is equivalent to TPR and sensitivity is the proportion of real credit card transactions predicted correctly by the model as fraudulent cases. On the other hand, precision is the proportion of predicted observations such as fraudulent credit card transactions predicted by the model that are accurate [11]. If the recall is equal to 1, it indicates that all the credit card transactions are classified as fraudulent. Conversely, precision will be low as many non-fraudulent credit card transactions will be falsely classified as fraud. Thus, performance measurements such as the F1-measure give equal consideration to precision and recall. Moreover, the misclassification rate or error rate will be used which determines the percentage of misclassified observations by the model [30]. These measures were defined as follows.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{4}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{5}$$

$$\text{F} - \text{measure} = 2\,(\text{Precision} * \text{Recall})(\text{Precision} + \text{Recall}) \tag{6}$$

$$\text{Misclassification Rate} = (\text{FP} + \text{FN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{7}$$

False cases that are predicted as possible fraud are costly in fraud detection, as they are taken for further investigation. The precise detection of cases of fraud helps to avoid costs resulting from missing a fraudulent activity (Type-I error), which is usually greater than falsely alleging fraud (Type-II error). Therefore, a Type-I error and Type-II error were used. FP provides the total of nonfraudulent firms that are mistakenly labeled as fraudulent, whereas Type-II error (false negative) indicates the sum of nonfraudulent firms that are incorrectly labeled as fraudulent [45,46].

The experiments were carried out on a Windows 10 computer with an Intel Core i7—10750H CPU (2.60 GHz 6 cores) and 16 GB RAM, using the Jupyter Notebook environment in an Anaconda Navigator platform.

6. Results and Discussion

This section presents the results and discussion from our proposed approach and compares the performance of developed hybrid models to the state-of-the-art machine learning algorithms, namely LR, RF, DT, XGBOOST, SVM, NB, Adaboost and LGBM. The single algorithms were compared in terms of prediction performance using their AUROC score to find which ones perform the best in this dataset and therefore are most suited for use as the first algorithm for the proposed hybrid models. Figure 7 illustrates that

generally, all the single models (other than NB) gave relatively similar performance values (0.66–0.71). Adaboost achieved the highest score (0.71) in the first phase. The decision was made based on the highest TPR and lowest FPR achieved by Adaboost, while NB gives the worst performance with an AUROC score of 0.56. The low performance of NB relies completely on the independence assumptions, whereas the used dataset might have some dependence features. However, it demonstrated one of the highest performance rates in the AUC-PR (Figure 8) alongside SVM and LGBM. One the other hand, DT and LR has shown the worst performance with 0.22 and 0.28 AUC-PR measure, respectively.

As a result of the superior performance of Adaboost in terms of its AUCROC measurement, it was selected as the optimum single baseline model and was be combined with the rest of the algorithms to determine the best hybrid model. The Adaboost algorithm was able to correctly classify 9023 credit card transactions as shown in Table 3. Next, to establish the correctly classified dataset, TP was added to TN to train and validate the hybrid models.

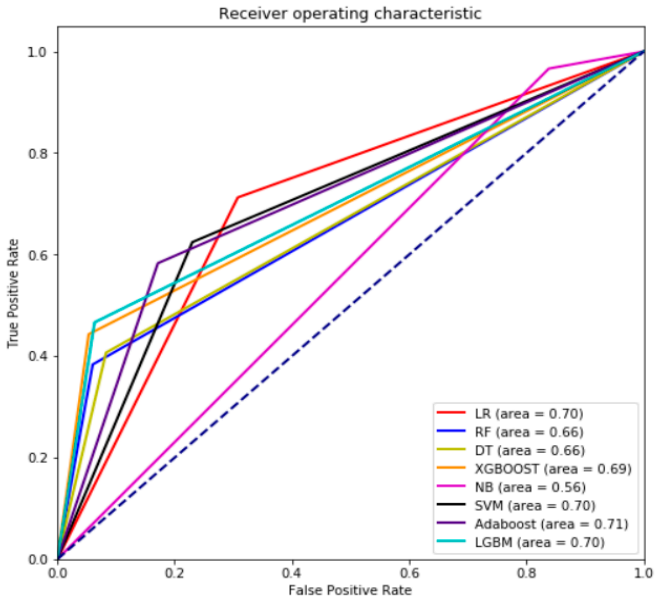


Figure 7. AUROC curve of the single models.

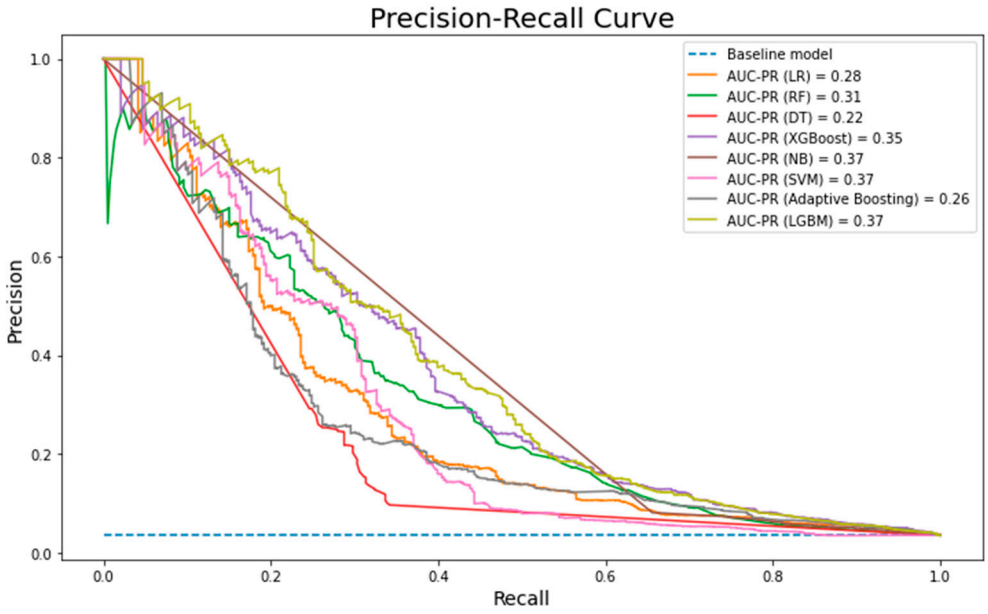


Figure 8. AUC-PR of the single models.

Table 3. Adaboost confusion matrix.

	Predicted Positive	**Predicted Negative**
Actual Positive	8798	1889
Actual Negative	161	225

In the second phase, seven hybrid machine learning models were developed (Figure 9). The predictive performance of the hybrid models has shown that the performance of the hybrid model Adaboost + LGBM excels in terms of its AUROC measure when utilizing real-world dataset (IEEE– CIS). As displayed in Table 4, the experimental results show that most of our proposed approaches outperformed the state-of-the-art machine learning algorithms in terms of AUROC, Type-I error, Type-II error, F1-measure, precision, misclassification rate and TNR, although some of the hybrid models (Adaboost +LR, Adaboost + NB and Adaboost + SVM) had a higher Type-II error than the state-of-the-art algorithms. However, this will not be a server issue as Type-I error is more costly and being able to lower such an error will have a good impact on the bank system. Additionally, all the proposed hybrid models were able to detect the non-fraudulent cases that were identified as non-fraud at a rate of almost 0.99 percent.

Table 4. Comparison table of the state-of-the-art machine learning algorithms and the proposed hybrid models.

	ROC	**Recall (TPR)**	**Precision**	**F-Measure**	**Misclassification Rate**	**TNR**	**Type-I Error**	**Type-II Error**
state-of-the-Art								
LR	0.70	0.71	0.08	0.14	0.30	0.68	0.31	0.28
RF	0.66	0.38	0.19	0.25	0.07	0.93	0.06	0.59
DT	0.66	0.41	0.15	0.22	0.10	0.91	0.08	0.57
XGBOOST	0.69	0.44	0.23	0.30	0.07	0.93	0.06	0.52
NB	0.56	0.97	0.04	0.08	0.81	0.15	0.84	0.03
SVM	0.70	0.62	0.09	0.16	0.23	0.74	0.25	0.35
Adaboost	0.71	0.58	0.11	0.18	0.17	0.82	0.17	0.41
LGBM	0.70	0.47	0.21	0.29	0.07	0.92	0.07	0.52
Hybrid Models								
Adaboost+LR	0.67	0.36	0.83	0.50	0.004	0.999	0.0004	0.52
Adaboost+RF	0.74	0.50	0.97	0.66	0.003	0.999	0.0004	0.33
Adaboost+DT	0.76	0.54	0.51	0.52	0.006	0.990	0.0099	0.29
Adaboost+XGBOOST	0.79	0.59	0.94	0.73	0.002	0.996	0.0031	0.31
Adaboost+NB	0.76	0.96	0.05	0.10	0.105	0.579	0.5791	0.05
Adaboost+SVM	0.58	0.18	0.91	0.30	0.005	1.0	0.0000	0.66
Adaboost+LGBM	0.82	0.64	0.97	0.77	0.002	0.998	0.0018	0.25

It is reflected in the AUC-PR (Figure 10) that the combination of Adaboost + XGBOOST outperforms the other six machine learning algorithms. Furthermore, Adaboost + XGBOOST and Adaboost + LGBM have a high capability of accurately identifying fraudulent activities as they have the lowest misclassification error rate (0.002) of AUROC. Utilizing Adaboost as a preprocessing step yields a cleaner dataset, which is expected to result in a more accurate and robust model that gives rise to a positive impact on the dataset via lowering the error rate for all the used algorithms. However, Adaboost + LGBM indicated a noticeable performance as it reached 0.82. On the contrary, LR and SVM showed decreasing performance for AUROC measures when hybridization with Adaboost took place. This indicates that hybridization between machine learning algorithms does not necessarily lead to higher performance. Additionally, looking into details for Adaboost + LGBM, a precision value of 0.97 indicates that when the model predicted a positive result it was correct 97 percent of the time and a recall value of 0.64 indicates that the model was able to identify 64 percent of all positive values correctly. In terms of the tradeoff between

both measures, an F1-measure of 77 percent gives an equal consideration of both values. Having such a high result compared with other hybrid models in terms of its precision, ROC, F-measure, and misclassification rate, we conclude that Adaboost + LGBM is the best hybrid model for the given dataset in this study.

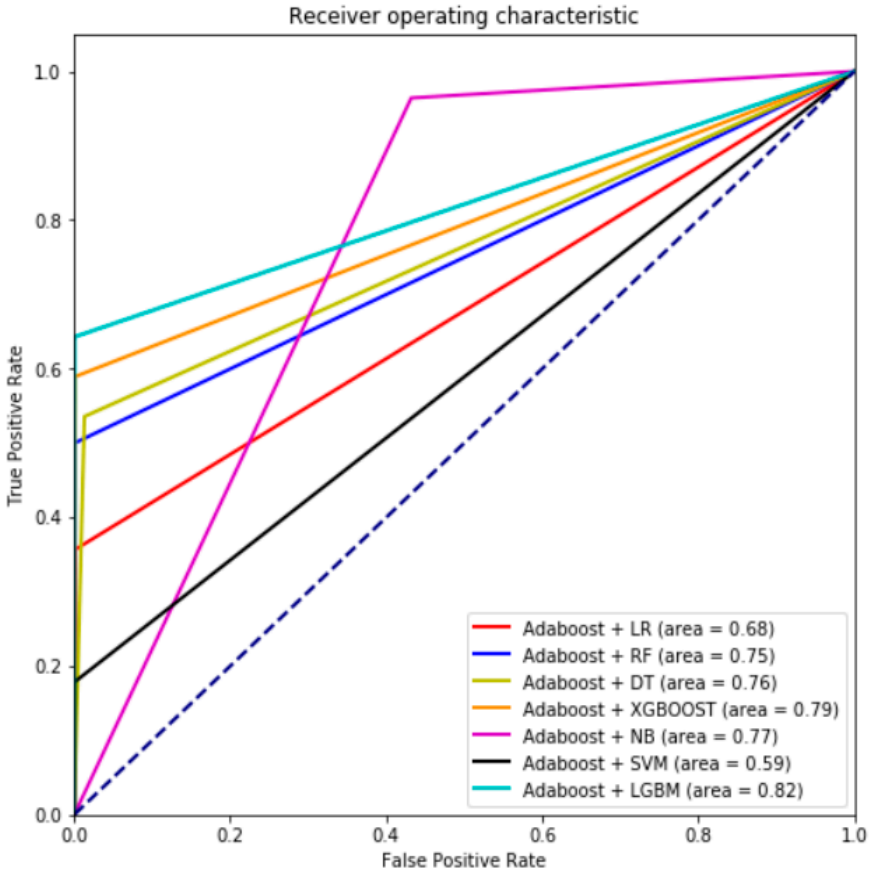


Figure 9. AUROC curve of the hybrid models.

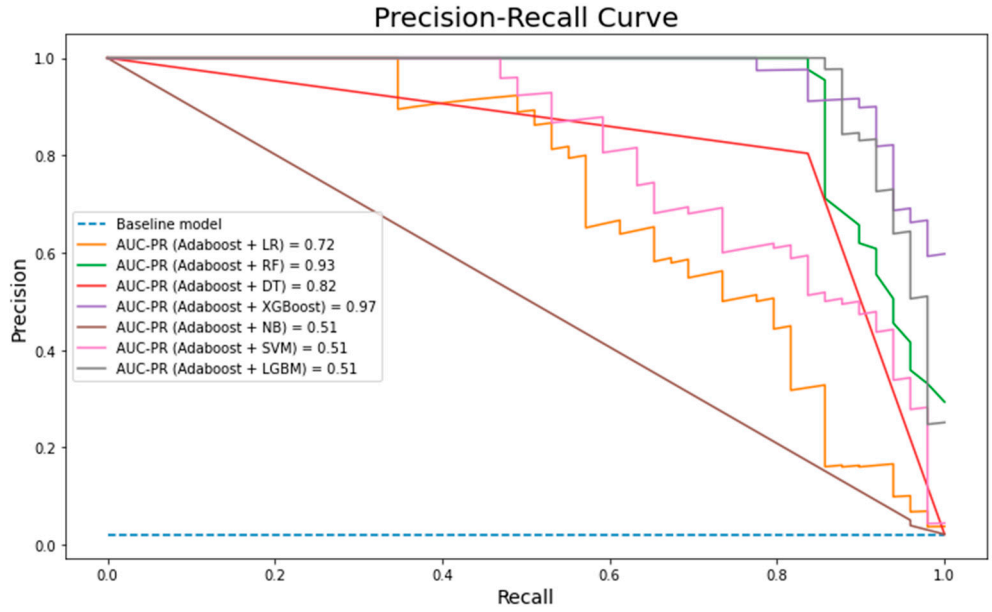


Figure 10. AUC-PR of the hybrid models.

7. Conclusions

Credit card fraud has recently become a major concern worldwide, especially for financial institutions. Various approaches have been previously used to detect fraudulent activities; however, the need to investigate different reliable methods still exists to detect fraudulent credit card transactions, as was the aim in this work for a single case study. In this research, several hybrid machine learning models were developed and investigated based on the combination of supervised machine learning techniques as a part of a credit card fraud detection study. The hybridization of different models was found to have the ability to yield a major advantage over the state-of-the-art models. However, not all hybrid models worked well with the given dataset. Several experiments need to be conducted to examine various types of models to define which works the best. Comparing the performance of the hybrid model to the state-of-the-art and itself, we conclude that Adaboost + LGBM is the champion model for this dataset. The result also illustrates that the use of hybrid methods has lowered the error rate. For future work, the hybrid models used in this study will be extended to other datasets in the credit card fraud detection domain.

Future work may focus on different areas, starting by proposing data preprocessing techniques to overcome the drawback of the missing values. Additionally, different methods of feature selection and extraction should be investigated in the credit card domain and to determine its impact on prediction accuracy. An investigation of the most appropriate hybrid model among the state-of-the-art machine learning algorithms to determine the most accurate hybridized model in the previously mentioned domain should be the main concern for future studies.

Author Contributions: Conceptualization, B.B. and X.C.; Data curation, E.F.M., B.B. and W.P.W.; Formal analysis, E.F.M. and K.W.K.; Funding acquisition, K.W.K. and X.C.; Investigation, E.F.M. and K.W.K.; Methodology, E.F.M., W.P.W. and X.C.; Project administration, K.W.K., B.B., W.P.W. and X.C.; Resources, B.B.; Supervision, X.C.; Writing—original draft, E.F.M.; Writing—review & editing, K.W.K. and W.P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Universiti Sains Malaysia, Short Term Grant [Grant Number: 304/PMGT/6315513], with the project entitled "The efficiency of the feature sampling interval scheme for the multivariate coefficient of variation in short production runs".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset analyzed during the current study is available in the Kaggle repository, https://www.kaggle.com/c/ieee-fraud-detection (accessed on 5 December 2021).

Acknowledgments: The authors are thankful to the School of Management and School of Computer Sciences, USM for providing us with the resources to conduct this research.

Conflicts of 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.

References

1. PWC. *Fighting Fraud: A Never-Ending Battle*; PWC: London, UK, 2020.
2. Garner, B.A. *Black's Law Dictionary, (Black's Law Dictionary (Standard Edition))*, 8th ed.; Thomson West: Toronto, ON, Canada, 2004; p. 1805.
3. Kültür, Y.; Çağlayan, M.U. Hybrid approaches for detecting credit card fraud. *Expert Syst.* **2017**, *34*, 1–13. [CrossRef]
4. Kurshan, E.; Shen, H. Graph Computing for Financial Crime and Fraud Detection: Trends, Challenges and Outlook. *Int. J. Semant. Comput.* **2020**, *14*, 565–589. [CrossRef]
5. West, J.; Bhattacharya, M. Intelligent Financial Fraud Detection: A Comprehensive Review. *Comput. Secur.* **2015**, *57*, 47–66. [CrossRef]
6. Ethem, A. *Introduction to Machine Learning*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2014.
7. Mater, A.C.; Coote, M.L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559. [CrossRef]

8. Hossain, M.A.; Islam, S.M.S.; Quinn, J.M.W.; Huq, F.; Moni, M.A. Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *J. Biomed. Inform.* **2019**, *100*, 103313. [CrossRef]
9. Abdelrahman, O.; Keikhosrokiani, P. Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning. *IEEE Access* **2020**, *8*, 189661–189672. [CrossRef]
10. Khan, M.A.; Ashraf, I.; Alhaisoni, M.; Damaševičius, R.; Scherer, R.; Rehman, A.; Bukhari, S.A.C. Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists. *Diagnostics* **2020**, *10*, 1–19. [CrossRef]
11. Cruz, J.A.; Wishart, D.S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2006**, *2*, 59–77. [CrossRef]
12. Lalmuanawma, S.; Hussain, J.; Chhakchhuak, L. Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals* **2020**, *139*, 110059. [CrossRef]
13. Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12*, 878. [CrossRef]
14. Taha, A.A.; Malebary, S.J. An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. *IEEE Access* **2020**, *8*, 25579–25587. [CrossRef]
15. Khandani, A.E.; Kim, A.J.; Lo, A.W. Consumer credit-risk models via machine-learning algorithms. *J. Bank. Financ.* **2010**, *34*, 2767–2787. [CrossRef]
16. Randhawa, K.; Loo, C.K.; Seera, M.; Lim, C.P.; Nandi, A.K. Credit Card Fraud Detection Using AdaBoost and Majority Voting. *IEEE Access* **2018**, *6*, 14277–14284. [CrossRef]
17. Krivko, M. A hybrid model for plastic card fraud detection systems. *Expert Syst. Appl.* **2010**, *37*, 6070–6076. [CrossRef]
18. Alharbi, A.; Alshammari, M.; Okon, O.D.; Alabrah, A.; Rauf, H.T.; Alyami, H.; Meraj, T. A Novel text2IMG Mechanism of Credit Card Fraud Detection: A Deep Learning Approach. *Electronics* **2022**, *11*, 756. [CrossRef]
19. Behera, T.K.; Panigrahi, S. Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network. In Proceedings of the 2015 2nd IEEE International Conference on Advances in Computing and Communication Engineering, Dehradun, India, 1–2 May 2015; pp. 494–499. [CrossRef]
20. Seeja, K.R.; Zareapoor, M. FraudMiner: A novel credit card fraud detection model based on frequent itemset mining. *Sci. World J.* **2014**, *2014*, 252797. [CrossRef]
21. Sarno, R.; Dewandono, R.D.; Ahmad, T.; Naufal, M.F. Hybrid Association Rule Learning and Process Mining for Fraud Detection. *IAENG Int. J. Comput. Sci.* **2015**, *42*, 59–72.
22. Carcillo, F.; Le Borgne, Y.A.; Caelen, O.; Kessaci, Y.; Oblé, F.; Bontempi, G. Combining unsupervised and supervised learning in credit card fraud detection. *Inf. Sci.* **2019**, *557*, 317–331. [CrossRef]
23. Li, S.H.; Yen, D.C.; Lu, W.H.; Wang, C. Identifying the signs of fraudulent accounts using data mining techniques. *Comput. Hum. Behav.* **2012**, *28*, 1002–1013. [CrossRef]
24. Sivanantham, S.; Dhinagar, S.R.; Kawin, P.A.; Amarnath, J. Hybrid Approach Using Machine Learning Techniques in Credit Card Fraud Detection. In *Advances in Smart System Technologies*; Springer: Singapore, 2021.
25. IEEE Computational Intelligence Society. IEEE-CIS Fraud Detection Can You Detect Fraud from Customer Transactions? 2019. Available online: https://www.kaggle.com/c/ieee-fraud-detection/overview (accessed on 5 December 2021).
26. Aoife, D.; Brian, M.; John, D.K. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*; The MIT Press: Cambridge, MA, USA, 2015.
27. Cerda, P.; Varoquaux, G.; Kégl, B. Similarity encoding for learning with dirty categorical variables. *Mach. Learn.* **2018**, *107*, 1477–1494. [CrossRef]
28. Qi, Z.; Zhang, Z. A hybrid cost-sensitive ensemble for heart disease prediction. *BMC Med. Inform. Decis. Mak.* **2020**, *21*, 73. [CrossRef]
29. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
30. Al Khaldy, M.; Kambhampati, C. Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset. *Int. Robot. Autom. J.* **2018**, *4*, 37–45. [CrossRef]
31. Lavanya, D.; Rani, D.K.U. Analysis of Feature Selection with Classification: Breast Cancer Datasets. *Indian J. Comput. Sci. Eng.* **2011**, *2*, 756–763.
32. Zhang, Y.; Wang, Z. Customer Transaction Fraud Detection Using Xgboost Model. In Proceedings of the 2020 International Conference on Computer Engineering and Application, Guangzhou, China, 18–20 March 2020; pp. 554–558. [CrossRef]
33. Sanz, H.; Valim, C.; Vegas, E.; Oller, J.M.; Reverter, F. SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinform.* **2018**, *19*, 1–18. [CrossRef]
34. Prati, R.C.; Batista, G.E.; Monard, M.-C. Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior. In Proceedings of the Mexican International Conference on Artificial Intelligence, Mexico City, Mexico, 26–30 April 2004; Springer: Berlin/Heidelberg, Germany, 2004.
35. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
36. Le, T.; Vo, M.T.; Vo, B.; Lee, M.Y.; Baik, S.W. A Hybrid Approach Using Oversampling Technique and Cost-Sensitive Learning for Bankruptcy Prediction. *Complexity* **2019**, *2019*, 8460934. [CrossRef]

37. Al-Hashedi, K.G.; Magalingam, P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Comput. Sci. Rev.* **2021**, *40*, 100402. [CrossRef]
38. Tsai, C.F.; Lin, W.C. Feature selection and ensemble learning techniques in one-class classifiers: An empirical study of two-class imbalanced datasets. *IEEE Access* **2021**, *9*, 13717–13726. [CrossRef]
39. Tsai, C.F.; Chen, M.L. Credit rating by hybrid machine learning techniques. *Appl. Soft Comput. J.* **2010**, *10*, 374–380. [CrossRef]
40. Bhattacharyya, S.; Jha, S.; Tharakunnel, K.; Westland, J.C. Data mining for credit card fraud: A comparative study. *Decis. Support Syst.* **2011**, *50*, 602–613. [CrossRef]
41. Vieira, S.; Pinaya, W.H.L.; Mechelli, A. *Introduction to Machine Learning*; MIT Press: Cambridge, MA, USA, 2019.
42. Harrington, P. *Machine Learning in Action*; Manning Publications, Co.: Shelter Island, NY, USA, 2012.
43. Faraji, Z. A Review of Machine Learning Applications for Credit Card Fraud Detection with A Case study. *J. Manag.* **2022**, *5*, 49–59. [CrossRef]
44. Lim, K.S.; Lee, L.H.; Sim, Y.-W. A Review of Machine Learning Algorithms for Fraud Detection in Credit Card Transaction. *IJCSNS Int. J. Comput. Sci. Netw. Secur.* **2021**, *21*, 31–40.
45. Hooda, N.; Bawa, S.; Rana, P.S. Fraudulent Firm Classification: A Case Study of an External Audit. *Appl. Artif. Intell.* **2018**, *32*, 48–64. [CrossRef]
46. Gepp, A.; Kumar, K.; Bhattacharya, S. Lifting the numbers game: Identifying key input variables and a best-performing model to detect financial statement fraud. *Account. Financ.* **2021**, *61*, 4601–4638. [CrossRef]