



An intelligent payment card fraud detection system

Manjeevan Seera¹ · Chee Peng Lim² · Ajay Kumar³ · Lalitha Dhamotharan⁴ · Kim Hua Tan⁵

Accepted: 3 June 2021 / Published online: 8 June 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Payment cards offer a simple and convenient method for making purchases. Owing to the increase in the usage of payment cards, especially in online purchases, fraud cases are on the rise. The rise creates financial risk and uncertainty, as in the commercial sector, it incurs billions of losses each year. However, real transaction records that can facilitate the development of effective predictive models for fraud detection are difficult to obtain, mainly because of issues related to confidentiality of customer information. In this paper, we apply a total of 13 statistical and machine learning models for payment card fraud detection using both publicly available and real transaction records. The results from both original features and aggregated features are analyzed and compared. A statistical hypothesis test is conducted to evaluate whether the aggregated features identified by a genetic algorithm can offer a better discriminative power, as compared with the original features, in fraud detection. The outcomes positively ascertain the effectiveness of using aggregated features for undertaking real-world payment card fraud detection problems.

Keywords Classification · Feature aggregation · Fraud detection · Payment card · Predictive modeling

1 Introduction

Many types of payment cards, which include credit, charge, debit, and prepaid cards, are widely available nowadays, and they constitute one of the most popular payment methods in some countries (Pavía et al., 2012). Indeed, advances in digital technologies have transformed the way we handle money. Payment methods have changed from being a physical activity to a digital transaction over electronics means (Pavía et al., 2012). This has revolutionized the landscape of monetary policy, including business strategies and operations of both large and small companies.

As reported in Forbes (2011) by the American Bankers, it is estimated that 10,000 transactions pertaining to payment cards occur every second globally. Owing to such a high transaction rate, payment cards have become a target for fraud. Fraud has been a key concern in most commercial and business areas (Bernard et al., 2019). Indeed, since

✉ Manjeevan Seera
manjeevansingh.seera@monash.edu

Extended author information available on the last page of the article

Diners Club issued the first credit card in 1950, credit card companies have been constantly fighting against fraud (Forbes, 2011). Each year, payment card fraud leads to losses in billions of dollars. These losses create risk and uncertainties to the financial institutions (Sariannidis et al., 2020). Fraud cases occur under different conditions, e.g., transactions at the Point of Sales (POS), online or over-the-telephone transactions, i.e., Card Not Present (CNP) cases, or transactions with lost or stolen cards. The loss from fraudulent incidents in payment cards amounted to \$21.84 billion in 2015, with issuers bearing the cost of \$15.72 billion (The Nilson Report, 2016). Based on the European Central Bank in 2012, the majority (60%) of fraud cases stemmed from CNP transactions, while another 23% at the POS terminals.

The potential of substantial monetary gains, combined with the ever-changing nature of financial services, creates a wide range of opportunities for fraud cases to occur (Ferreira & Meidutė-Kavaliauskienė, 2019). Funds from payment card fraud are often used in criminal activities, e.g., to support terrorism activities (Everett, 2003). Over the years, fraudulent mechanisms have evolved along with the models used by the banks to avoid fraud detection (Bhattacharyya et al., 2011). Therefore, it is imperative to develop effective and efficient methods to detect payment card fraud. The developed methods also need to be revised continually in accordance with the advances in technologies. There are many challenges in developing effective fraud detection methods. Among them, researchers face the difficulty in obtaining real data samples from payment card transactions as financial institutions are reluctant to share their data owing to confidentiality issues (Dal Pozzolo et al., 2014). As a result, only limited research studies with real data are available in this area. Some machine learning and related studies on fraud detection have been conducted using publicly available data sets (Sahin et al., 2013), e.g. SVMs (support vector machines), ANNs (artificial neural networks), decision trees, as well as regression and rule-based methods.

In this paper, a total of thirteen widely used statistical and machine learning methods for fraud detection in payment cards are implemented. The methods used include SVMs and ANNs, as well as more recent deep learning methodologies. Two data sources are used for evaluation: publicly available repositories and real payment card database. In this study, the availability of real payment card data for evaluation is of particular importance, which ensure that the methods developed are usable and useful in real-world situations. In other words, the developed methods should be deployable in the financial sector for detecting payment card fraud, leading to reduction in financial losses and mitigation of risk and uncertainty in the business world.

The main contributions of the paper are two-fold. Firstly, a feature aggregation method is devised for processing payment transaction records. The key benefit of aggregating various features from transaction data is improvement in robustness to counter the effects of concept drift (Whitrow et al., 2009). Besides that, feature selection using optimization methods is applied to the transaction data. Secondly, a real payment card database is used for analyzing the performance of a variety of statistical and intelligent data-based algorithms for fraud detection, in addition to using benchmark data. As it is different to obtain real financial records for analysis (due to confidentiality issues), the outcome of our study is important for deriving valuable insights into the robustness of various data-based methods utilizing aggregated features for fraud detection in payment card transactions in real-world environments.

For the remaining part of this paper, we firstly present a literature review on finance applications of statistical and machine learning algorithms. Secondly, the background of various classification methods devised in this paper is given. We then explain the detailed experimental study, which covers the results, analysis, and comparison. Implications of the

developed methods for practical application and a summary of the findings are presented at the end of this paper.

2 Literature review

Fraud detection systems are used in identifying unusual behaviors in electronic payment transactions. In this paper, we conduct a review on a variety of fraud detection systems, which is divided into two broad categories, namely, systems that use the original features and those that aggregate the original features. A summary of the review is presented at the end of this section.

2.1 Original features

A data set from a Brazilian online payment service was used in de Sá et al. (2018) for detection of credit card fraud. A customized Bayesian Network Classifier was proposed. The underlying method relied on an evolutionary algorithm coupled with a hyper-heuristic mechanism to search for the best component combinations in the data set. The proposed algorithm improved the efficiency by 72.64% (de Sá et al., 2018). A data set from Worldline Belgium was used in Van Vlasselaer et al. (2015) for credit card fraud identification with respect to the online stores sector. A method known as APATE (Anomaly Prevention using Advanced Transaction Exploration) that combined customer spending history and a time-based suspiciousness measure pertaining to each transaction was formulated, which yielded good results in term of the area under the ROC (Receiver Operating Characteristic) (i.e. AUC) plot (Van Vlasselaer et al., 2015).

In Russac et al. (2018), the same data set as in Van Vlasselaer et al. (2015) was used to extract sequential information from the transactions. As the computational load was heavy, only three categorical features were used. Sequential information of the three categorical features was extracted using a Word2Vec neural network. This method led to reduction of the memory usage by half while improving the performance by 3% (Russac et al., 2018). A real data set from Banco Bilbao Vizcaya Argentaria (BBVA) was used in Gómez et al. (2018) for fraud detection. The MLP (multi-layer perceptron) ANN was used in classifying the data samples. Various backpropagation methods were used together with the MLP network. The results were comparable with those from other costly solutions (Gómez et al., 2018). A databased containing transaction records of credit cards was used for fraud detection in Jurgovsky et al. (2018). The LSTM (long short term memory) model was applied to examine the sequence of transactions. The RF (Random Forest) model was compared with LSTM in the study. Based on the observation, both RF and LSTM detected different fraud cases. A further analysis suggested a combination of RF and LSTM could result in a better fraud detection system (Jurgovsky et al., 2018).

A data set from a card processing company, CardCom, was used in Robinson and Aria (2018) to ascertain fraud cases in prepaid cards. The Hidden Markov Model (HMM) was exploited. The proposed technique acquired a good F-score, which could detect fraudulent cases in real-time (Robinson & Aria, 2018) with thousands of prepaid card transactions. A total of three real-world data sets were used in Rtayli and Enneya (2020). A hybrid SVM model consisting of recursive feature elimination, grid search and oversampling technique was proposed. The proposed method gave the best results in terms of efficiency and effectiveness (Rtayli & Enneya, 2020). The weighted extreme learning machine (WELM) was

used for evaluation with benchmark credit card data in Zhu et al. (2020). WELM with the dandelion algorithm yielded a high detection performance (Zhu et al, 2020). An ensemble model using sequential modeling of deep learning and voting mechanism of ANN was proposed in Forough and Momtazi (2021). Based on real-world credit card data, the time analysis results indicated the real time high efficiency of the proposed model as compared with other models (Forough & Momtazi, 2021).

2.2 Aggregated features

In this section, we review feature aggregation for fraud detection. Among various feature aggregation methods, feature averaging summarizes the cardholder activities by comparing the spending habits and patterns (Russac et al., 2018). In Bahnsen et al. (2016), a credit card-related database for fraud detection was examined. By analyzing the periodic behaviors over time, an aggregated feature set was generated. The use of aggregated features yielded an average of 13% saving with different classification models, including RF (random forest), LOR (logistic regression), and DT (decision tree) (Bahnsen et al., 2016).

Two data sets pertaining to European card holders were used in Dal Pozzolo et al. (2017) for detecting credit card fraud. An alert-feedback interaction was applied to train and update the classifier. During augmentation of the features, a set of aggregated features linked with individual transactions was generated to better separate fraud cases from real transactions. The outcome indicated that a lower degree of influence of feedback led to less precise alerts (Dal Pozzolo et al., 2017). In Fu et al. (2016), a credit card database collected from a commercial bank was collected for fraud detection analysis. To capture the underlying patterns associated with fraudulent behaviours, the CNN (convolutional neural network) was employed. Feature engineering techniques similar to transaction aggregation were adopted to generate a total of eight features, from average transaction amount to trading entropy gain. Given a feature matrix, the CNN was able to identify the patterns of fraud and produced better performance when compared with other methods (Fu et al., 2016).

In Jiang et al. (2018), a simulator was used to generate credit card transaction data for the purpose of fraud detection. All generated data samples were divided into multiple groups, with each group containing similar transactional behaviors. A window-sliding method was applied to aggregate the transactions and produce 7 (4 amount-related and 3 time-related) features. The RF achieved 80% accuracy at the detection of transactions, coupled with a feedback mechanism (Jiang et al., 2018). Another fraud detection study pertaining to credit cards with simulated data was reported in Lim et al. (2014). A conditional weighted transaction aggregation method was applied to record the transactions. Each transaction was given a weight. A distance measure between the previous and current transactions was exploited to set the weight. Algorithms such as RF, k -NN, and LOR were used for classification. Comparing with the transaction-based technique, the method of aggregation was able to produce better outcomes (Lim et al., 2014).

In Lucas et al. (2020), modelling on the sequence of credit card transactions was conducted. Three different scenarios were considered, namely sequences fraudulent and non-fraudulent records, sequences obtained by fixing cardholders, and sequences of amount spent between the current and past transaction records. Each sequence was associated with a likelihood, as modelled using the HMM model. The resulting information was adopted as additional features for analysis with the RF model. The results indicated that the feature engineering performed well for credit card fraud detection (Lucas et al., 2020). A real-world credit card data set from a commercial bank in China was used by Zhang et al.

(2020). A method that consisted of feature engineering and deep learning was devised. For each occurrence of a transaction, a number of feature variables were computed based on the incoming and past transaction records for aggregation. In addition, homogeneous historical transactions were considered. The proposed method could efficiently identify fraudulent transactions (Zhang et al., 2020).

2.3 Remarks

Table 1 depicts a summary of the relevant publications discussed earlier. In general, two broad categories can be formed: original features and aggregated features. According to the review, two types of features are typically used, i.e., original/standard and aggregated features. The use of aggregated features increases the ability of the classifiers to detect fraudulent transactions, as depicted by the reported results. On the other hand, there is a lack of comprehensive statistical analyses for the results presented in the literature. In general, the accuracy metric is used for performance assessment, which could not aptly reflect the true potential of a classifier, especially in dealing with highly imbalanced data sets that commonly exist in fraud detection studies. Some studies use the AUC metric, which is a more comprehensive performance indicator. One critical issue is the false alarm rate, i.e., genuine transactions flagged by the detection systems as fraudulent cases. Financial institutions spend substantial time and money in investigating these legitimate cases. Importantly, flagging a genuine transaction causes customer dissatisfaction and inconvenience. As such, an efficient and effective fraud detection system should minimise the false alarm rate, which is the main focus of our study in this paper. Our main contribution is we deploy a real-world payment data set to comprehensively assess the ability of a total of thirteen classifiers along with the use of statistical measures, including the AUC, for performance assessment as well as for comparison with other results reported in the literature.

3 Classification methods

In this section, we present an overview on thirteen classification methods devised for evaluation. These methods are split into six main sub-groups, with the details as follows.

3.1 Bayesian methods

The Bayesian theorem is used to develop the Naïve Bayes (NB) method. To formulate a classification method, an independence assumption is adopted in NB. Features from different classes are assumed to be independent from each other, which is a strong assumption. Given an input vector \mathbf{X} , with Y as the associated target class, the Bayesian theorem yields

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{P(\mathbf{X})}, \quad (1)$$

where $P(\mathbf{X}|Y)$ and $P(Y|\mathbf{X})$ are the conditional probability (of \mathbf{X} given the occurrence of Y) and posterior probability (of Y given the occurrence of \mathbf{X}), respectively; while $P(\mathbf{X})$ and $P(Y)$ are the probability of evidence with respect to \mathbf{X} and Y . The predicted target class of input \mathbf{X} is based on $P(Y|\mathbf{X})$ that yields the highest value. Suppose that input \mathbf{X} consists of n features, we have

Table 1 Summary of review

Features	References	Data set	Classifier	Remarks	
Original	de Sá et al. (2018)	Brazilian company	Bayesian	Improved efficiency by 72.64%	
	Van Vlasselaer et al. (2015)	Worldline Belgium	APATE	Best AUC acquired with addition of customer spending history	
	Russac et al. (2018)		Word2Vec	Performance improved by 3%	
	Gómez et al. (2018)	BBVA bank	MLP	Solution comparable with costly ones	
	Jurgovsky et al. (2018)	Credit card data	RF, LSTM	RF + LSTM could result in a better fraud detection system	
	Robinson and Aria (2018)	CardCom	HMM	Able to detect fraudulent cases in real-time	
	Rayli and Enneya (2020)	Credit card data	Hybrid SVM	Recursive feature elimination and hyper-parameters optimization	
	Zhu et al. (2020)	Benchmark data	WELM	Dandelion algorithm with probability-based mutation outperforms particle swarm optimization	
	Aggregated	Forough and Momiati (2021)	Credit card data	Deep learning	Efficient real-time performance
		Bahnsen et al. (2016)	European bank	DT, LOR, RF	Average saving of 13% was achieved
Dal Pozzolo et al. (2017)		European bank	Ensemble	Lower degree of influence of feedback led to less precise alerts	
Fu et al. (2016)		Commercial bank	CNN	Able to identify the patterns of fraud and produce better performance	
Jiang et al. (2018)		Simulated data	RF	80% detection accuracy	
Lim et al. (2014)			RF, k-NN, LOR	Aggregation-based methods outperformed transaction-based methods	
Lucas et al. (2020)		Credit card data	HMM	Useful feature engineering for classification	
Zhang et al. (2020)		Commercial bank in China	HOBA	Aggregation based on previous and incoming transactions	

$$P(Y|\mathbf{X}) = \frac{P(Y) * \prod_{i=1}^n P(\mathbf{X}_i|Y)}{P(\mathbf{X})}, \quad (2)$$

3.2 Tree methods

A total of five methods based on trees are presented, which consists of the Decision Trees (DT), Decision Stump (DS), Random Tree (RT), Random Forest (RF) and Gradient Boosted Trees (GBT).

Decision Trees (DT) can be applied to both classification and regression problems. In classification tasks, a predicted outcome with respect to a target class is required, while the predicted outcome can use a real number (e.g. the price of a stock) in regression tasks. A DT is used for predicting dependent Y from independent variables $X=X_1, \dots, X_n$. In DT, a number of nodes are established, forming a link from the input to output data samples. The Gini impurity measure is used to determine how frequently a randomly chosen input sample is incorrectly labeled.

The Gini index for a data set with J classes can be computed using

$$I_G = 1 - \sum_{k=1}^J p_k^2 \quad (3)$$

where $i \in \{1, 2, \dots, J\}$, and p_k represents the proportion of samples in class k . A rule that splits the input features is encoded in each DT node. To form a tree structure, new nodes can be created, subject to a stopping criterion. When an input sample is provided, the majority of data samples stemmed from a leaf node of the tree are identified, leading to a predicted target class.

On the other hand, Decision Stump (DS) is developed based on DT with only single split. DS is useful to tackle an uneven distribution of data samples. The procedure of tree averaging uses a set of n weights, $\{w_1, w_2, \dots, w_n\}$, for every tree in the set of pruned trees. The weights are normalized, leading to $\sum_{i=1}^{i=n} w_i = 1$. Object O can be classified by a smoothed tree, in which the probability of each class given by each tree is determined as follows

$$Pr(O_i^k) = Prob(O \text{ belongs to } k) \quad (4)$$

where k is the class. Then, calculate the distribution probability over all classes with respect to object O by adding over the pruned trees set,

$$Prob(O \text{ belongs to } k) = \sum_{i=1}^{i=n} w_i \times Pr(O_i^k) \quad (5)$$

The Random Tree (RT) method is devised based on the same principle of DT. It, however, uses a subset of randomly selected input samples for the split process. A subset ratio is exploited to determine the random subset size. Both nominal and numerical data samples can be used with an RT. Each interior node corresponds to one of the input variables. The number of edges in an interior node is equivalent to the possible values pertaining to the corresponding input variable.

Based on an ensemble of RT models, the Random Forest (RF) method is formed, where the tree size is a user-defined parameter. Given the predictions from the trees, the final

predicted class yielded by RF is determined by using a voting mechanism. When each classifier, denoted as $h_k(\mathbf{x})$, is a DT, the ensemble is an RF. The DT parameter of $h_k(\mathbf{x})$ is

$$\Theta_k = (\theta_{k1}, \theta_{k2}, \dots, \theta_{kp}) \quad (6)$$

The term can be written as $h_k(\mathbf{x}) = h(\mathbf{x}|\Theta_k)$. For classification function $f(\mathbf{x})$ combines all classifier outputs, whereby every DT outputs a vote for the most probable class given input \mathbf{x} . The class with the highest number of votes wins.

Another useful regression/classification ensemble method is Gradient Boosted Trees (GBT). To increase the prediction predictions, GBT exploits a forward-learning ensemble with the boosting concept for improvement. Using a training set of $\{(x_1, y_1), \dots, (x_n, y_n)\}$ of x inputs and the corresponding y outputs, the goal of GBT is to find $\hat{F}(x)$ to function $F(x)$ which reduces the expected value of loss function $L(y, F(x))$, i.e.,

$$\hat{F} = \arg \min_F E_{x,y} [L(y, F(x))] \quad (7)$$

GBT assumes y seeks an approximation of $\hat{F}(x)$ in form of a weighted sum of functions $h_i(x)$ from some class H (known as weak learners),

$$\hat{F}(x) = \sum_{i=1}^M y_i h_i(x) + \text{const} \quad (8)$$

3.3 Neural networks

In general, the Multi-Layer Perceptron (MLP) Artificial Neural Network (ANN) structure comprises three (input, hidden, and output) layers. The backpropagation (BP) algorithm for learning the network weights. Similarly, BP is used for learning in the FF (Feed-Forward) ANN. In both networks, the information is propagated from the input to output nodes via the nodes in the hidden layer. The goal in training is to minimize the cost function C of data set D_p via

$$C(D_p, W) = \sum_{p=1}^P L(y(p), \hat{y}(p)) \quad (9)$$

where $y(p)$ is target (output) vector for p , \hat{y} is actual output of MLP ($\hat{y} = MLP(x; w)$), $x(p)$ is input vector for p , L is criterion to optimize the mean squared error. Gradient descent is used, which is an iterative procedure to modify the weights, W^t ,

$$W^{t+1} = W^t - n \frac{\partial C(D, W^t)}{\partial W^t} \quad (10)$$

where n is the learning rate.

Deep Learning (DL) has its roots in ANN. It is a popular learning model currently, where an FF ANN model with possibly thousands of hidden layers is formulated. Different activation functions are used in DL. Based on local data, the weights in each node makes contributions toward a global model through an averaging procedure. As DL is based on a multi-layered ANN model, it is trained using SGD (stochastic gradient descent) with backpropagation. Backpropagation works by accumulating prediction errors from the forward phase for updating the network parameters so that it performs better in the next iteration.

The descent method needs the gradient loss function of $\nabla J(\theta)$ to be computed. In an ANN, the operation can be conducted by a computational graph. The graph turns the loss function J into a number of intermediate variables. Backpropagation is applied recursively in the chain-rule for computing the gradients from the outputs to the inputs.

3.4 Regression methods

Two different regression methods are presented, consisting of Linear Regression (LIR) and Logistic Regression (LOR). Given a data set, a linear function is used to regulate the relation with respect to the scalar variables in LIR. Specifically, linear prediction functions are exploited to model the relation based on parameters estimated using the data samples. When there are two or more predictors, the target output is a linear combination of the predictors. The output (dependent variable) is obtained using

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n, \quad (11)$$

where each b_i is the corresponding coefficient of each x_i , which is the explanatory variable. In a two-dimensional example, a straight line through the data samples is formed, whereby the predicted output, \hat{y} , for a scalar input x is given by

$$\hat{y} = b_0 + b_1x, \quad (12)$$

LOR, another regression method, is useful to process numerical as well as nominal input samples. In LOR, one or more predictors are exploited to yield an estimation of the probability pertaining to a binary response. Support the probability of event occurrence as p , the linear function of predictor x is given by

$$\text{logit} = \frac{\log p}{1 - p} = b_0x + b_1, \quad (13)$$

Similar to Eq. (13), in the case involving independent variables, x_i 's,

$$\text{logit} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n, \quad (14)$$

The output probability is computed using

$$p = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}}, \quad (15)$$

3.5 Support vector machines

Support Vector Machines (SVMs) are supervised models that can be used for both classification and regression tasks. A SVM can be formulated to act as a binary classifier through a model that assigns the data samples into two different categories. A parallel margin is established based on the data samples that yields the widest possible separation pertaining to different classes. Specifically, a hyperplane is established as follows

$$H = b + w \cdot x = 0, \quad (16)$$

where w , x , and b are the weight, input, and bias term, respectively. In an optimal hyperplane, H_0 , the margin, M , is given by

$$M = \frac{2}{\sqrt{w_0} \cdot w_0}, \quad (17)$$

where w_0 is formed from the training samples, known as the support vectors, i.e.,

$$|w_0 = \sum y_{ii}x_i|, \quad (18)$$

3.6 Rule induction

Rule Induction (RI) is a method where formal rules are extracted from a set of observations. The rules represent a full model or local patterns pertaining to the data sample. The method begins with less common classes, and grows as well as prunes the rules based on all positive samples, or over 50% of error is reached. For each rule, r_i , its accuracy is calculated using

$$A(r_i) = \frac{\text{Correct data covered by rule}}{\text{All data covered by rule}}, \quad (19)$$

During the growing phase, specific conditions are incorporated into the rule, and this process continues until an accuracy rate of 100% is achieved. During the pruning phase, the final sequence of each rule is removed using a pruning metric.

4 Empirical evaluation

In this section, an empirical evaluation using publicly available databases together with a real-world payment card database is presented. All experiments were performed using a commercial data mining software, i.e., RapidMiner Studio 7.6. RapidMiner Studio is a software product that allows the prototyping and validation of machine learning and related classification algorithms. All parameters were set based on the default settings in RapidMiner. For evaluation, we adopt the tenfold cross-validation (CV) method. The key advantage of CV is to minimize bias with respect to random sampling during the test phase (Seera et al., 2015).

4.1 Class imbalance

In an imbalanced data set, there are fewer training samples from the minority class(es), as compared with those from the majority class(es) (Rtayli & Enneya, 2020). In fraud detection, the number of fraud cases is normally tiny, as compared with those of normal transactions. Fraudsters always attempt to create a fraudulent transaction as close as possible to a real transaction, in order to avoid being detected (Li et al., 2021). This data imbalanced issue affects the performance of machine learning methods. The learning algorithms normally focus on data samples from the majority class, leading to a higher accuracy rate as compared with that of the minority class.

While there are a number of ways for tackling the class imbalance challenge, creating synthetic data samples for learning leads to an increase in the false positive rate (Fiore et al., 2019). In our evaluation, we adopt the under-sampling technique to tackle this

Table 2 Distribution of data samples

Data set	Features	Class 1	Class 2	Total
Statlog (German credit)	24	300	700	1000
Statlog (Australian credit)	14	307	383	690
Default of credit card	23	6636	23,364	30,000

Table 3 Accuracy results (best in bold)

Model	German (%)	Australia (%)	Card (%)
NB	72.700	80.000	70.700
DT	70.000	83.188	81.973
RF	70.000	81.304	77.950
GBT	74.000	86.232	82.060
DS	69.900	85.507	81.960
RT	70.000	70.725	79.153
ANN	70.200	83.188	81.827
MLP	73.900	85.652	81.963
LIR	76.700	85.797	79.920
LOR	76.500	86.087	81.050
SVM	75.700	85.362	80.863
RI	71.900	85.072	78.740
DL	72.600	85.217	81.737

imbalanced data distribution problem. It is desirable to avoid the mistake in classifying legitimate transactions as fraud cases, in order to avoid poor customer services. On the other hand, it is necessary to accurately detect fraudulent transactions, in order to minimize financial losses.

The use of accuracy in imbalanced classification problems is inappropriate (Han et al., 2011), because the majority of data samples influence the results. The ROC curve depicts the classification performance subject to different thresholds. The AUC score is the probabilities of a classifier's ranking pertaining to randomly selected positive samples to be larger than those of negative ones, which statistically significant difference in performances with respect to different methods (Hanley & McNeil, 1982). In our experiments, the AUC score was used.

4.2 Benchmark data

We used three data sets from the UCI Machine Learning Repository (Dua & Graff, 2019), namely Statlog (German Credit), Statlog (Australian Credit), and Default Credit Card; hereafter denoted as German, Australian, and Card, in the first evaluation. All three were binary classification tasks, and the number of features varied from 14 to 24. The details are given in Table 2.

The accuracy rates for the three problems are listed in Table 3. With respect to the German data set, the lowest (69.9%) and highest (76.7%) accuracy scores were produced by LIR and DS, respectively. For the Australia data set, the lowest accuracy rate was produced by RT (i.e., 70.725%), while GBT produced the highest accuracy rate (i.e., 86.232%). This

Table 4 AUC results (best in bold)

Model	German	Australia	Card
NB	0.763	0.901	0.736
DT	0.500	0.860	0.643
RF	0.590	0.908	0.678
GBT	0.767	0.937	0.778
DS	0.500	0.862	0.644
RT	0.506	0.682	0.545
ANN	0.725	0.905	0.745
MLP	0.777	0.926	0.742
LIR	0.796	0.932	0.717
LOR	0.789	0.934	0.723
SVM	0.786	0.929	0.709
RI	0.653	0.878	0.662
DL	0.770	0.934	0.772

Table 5 Comparison of accuracy and AUC using the German data set (best in bold)

Model	Accuracy (%)	AUC
DT (Feng et al., 2018)	67.00	0.610
NN (Feng et al., 2018)	70.00	0.620
SVM (Feng et al., 2018)	71.50	0.550
BagDT (Feng et al., 2018)	73.20	0.622
BagNN (Feng et al., 2018)	76.00	0.672
BagSVM (Feng et al., 2018)	75.00	0.651
RF (Feng et al., 2018)	74.00	0.635
<i>k</i> -NN (Jadhav et al., 2018)	75.20	0.759
NB (Jadhav et al., 2018)	73.70	0.767
LIR	76.70	0.796

constituted the most significant variation in all classifiers in this experiment, i.e., a difference of 16% from the lowest to the highest accuracy rates. For the Card data set, NB and GBT yielded the lowest and highest accuracy rates, i.e., 70.7% and 82.06%, respectively.

The AUC scores are given in Table 4. A score as close to unity as possible is preferred. In the German data set, the lowest AUC score of 0.5 was produced by both DT and DS, while LIR achieved the highest score of 0.796. RT and GBT yielded the lowest (0.682) and highest (0.937) AUC results for the Australia data set, respectively. Similarly, RT and GBT produced the lowest (0.545) and highest (0.778) AUC results for the Card data set, respectively.

To further demonstrate the usefulness of various methods, the results were compared with those in recent publications, as reported in Feng et al. (2018) and Jadhav et al. (2018). A fivefold CV method was used in Feng et al. (2018), while Jadhav et al. (2018) used the tenfold CV method, i.e., the same as our experiment. The models in both Feng et al. (2018) and Jadhav et al. (2018) were built using the MATLAB software. The highest accuracy rate of 76.70% in the German data set, as shown in Table 5, was achieved from our experiment. With respect to AUC, the best score of 0.796 was also from LIR in our study, followed by NB (Jadhav et al., 2018) at 0.767.

Table 6 Comparison of accuracy and AUC using the Australian data set (best in bold)

Model	Accuracy (%)	AUC
DT (Feng et al., 2018)	82.10	0.820
NN (Feng et al., 2018)	85.30	0.855
SVM (Feng et al., 2018)	85.40	0.860
BagDT (Feng et al., 2018)	85.50	0.860
BagNN (Feng et al., 2018)	86.00	0.860
BagSVM (Feng et al., 2018)	85.50	0.870
RF (Feng et al., 2018)	87.00	0.860
<i>k</i> -NN (Jadhav et al., 2018)	85.70	0.878
NB (Jadhav et al., 2018)	80.43	0.913
GBT	86.23	0.937

Table 7 Comparison of accuracy and AUC using the Card data set (best in bold)

Model	Accuracy (%)	AUC
DT (Feng et al., 2018)	82.00	0.665
NN (Feng et al., 2018)	82.05	0.660
SVM (Feng et al., 2018)	82.00	0.643
BagDT (Feng et al., 2018)	82.00	0.665
BagNN (Feng et al., 2018)	82.00	0.660
BagSVM (Feng et al., 2018)	81.00	0.620
RF (Feng et al., 2018)	82.00	0.625
<i>k</i> -NN (Jadhav et al., 2018)	80.80	0.627
NB (Jadhav et al., 2018)	71.36	0.699
GBT	82.06	0.778

In the Australian data set, the best accuracy rate of 87.00%, as shown in Table 6, was yielded by RF (Feng et al., 2018). In comparison with our results, the accuracy rate achieved by GBT was 86.23%. In addition, the best AUC result (0.937) was produced by GBT in our experiment, which was followed by 0.913 from NB (Jadhav et al., 2018). As can be observed in Table 7, the best reported accuracy rate with respect to the Card data set was 82.06% from GBT in our study. Similarly, GBT achieved the best AUC score of 0.778, which was higher than the best score of 0.699 from NB in (Jadhav et al., 2018). As can be observed in Table 7, the best reported accuracy rate with respect to the Card data set was 82.06% from GBT in our study. Similarly, GBT achieved the best AUC score of 0.778, which was higher than the best score of 0.699 from NB in (Jadhav et al., 2018).

4.3 Real payment card data

In this evaluation, we established a database with real payment card transactions obtained from a financial firm in Malaysia. The data set contained transactions from January to March 2017. A total of 61,786 transaction records were available for evaluation. The transactions covered activities in 112 countries, with various spending items ranging from online website purchases to grocery shopping. Among the transactions, 46% occurred locally in Malaysia, which was followed by transactions made in Thailand and Indonesia.

A total of 31 transactions were identified and labeled as fraud cases, with the remaining being genuine, or non-fraud cases. The list of features used is given in Table 8.

4.4 Feature aggregation

In feature engineering, feature selection and feature aggregation constitute two main considerations. The capability of extracting discriminative features and removing irrelevant ones is important to improve classification performances, especially when dealing with high-dimensional data (Rtayli & Enneya, 2020). In general, feature selection can be categorized into two: filter and wrapper methods (Zhang et al., 2019). An independent evaluation is used in filter-based method to evaluate and identify important features. A pre-determined classifier is used in computing the evaluation in wrapper-based methods, which is computationally expensive.

When the number of features is small, feature aggregation methods are useful. A number of aggregation methods used in the literature are reviewed. Specifically, the data set used in Bahnsen et al. (2016) contained 27 features. Eight groups of duration were established, namely 1, 3, 6, 12, 18, 24, 72, and 168 h. The features were aggregated based on merchant codes, merchant groups, various transaction types, modes of entry, and country groups. The raw features yielded inferior results than those of aggregated features in all cases (Bahnsen et al., 2016). In Jha et al. (2012), additional features were derived by combining information across 1 day, 1 month, and 3 months. A total of 14 derived features were created, covering transaction amounts over a month, the total transaction records in one country over the last month, and average spending amounts during the last 3 months.

In Whitrow et al. (2009), aggregation of transaction information over periods of 1, 3, and 7 days was conducted. The data set was grouped into 24 categories, leading to a total of 48 features using the Laplace-smoothed estimation of the fraud rates. The aggregated

Table 8 List of features

Features	D1	D2	D3	D4
Account no	✓	✓	✓	✓
Transaction amount	✓	✓	✓	
Transaction date	✓	✓	✓	✓
Transaction time	✓	✓	✓	✓
Device type	✓	✓	✓	
MCC	✓			
Acquiring country	✓	✓	✓	
For country	✓			
Transaction type	✓			✓
Transaction amount no		✓	✓	✓
Transaction amount sum		✓	✓	✓
Acquiring country no		✓	✓	✓
Acquiring country sum		✓	✓	✓
MCC no			✓	✓
MCC sum			✓	✓
Device type no			✓	✓
Device type sum			✓	✓

records included the numbers of transactions performed at terminals using PIN as well as the numbers of transactions performed to date. In Fu et al. (2016), a total of 8 features were used for aggregation, with no details of the time period. The features used for aggregation included average, total, bias, number, country, terminal, merchant, trading entropy of transactions during a period of time. In Jiang et al. (2018), a total of 5 raw features were used, and the users were divided into 3 similar groups using *k*-means. To combine the information from the transactions information, the sliding-window method was used. Aggregation over a period of a week, a month, or a user-defined period was conducted. A total of 7 features were aggregated, with 4 amount-related features and another 3 time-related features. Based on the original nine features in our real data set provided by a Malaysian financial institution, we investigated on how aggregation of transaction records could affect the classifiers' performance. Information on each cardholder's account was continuously updated as new transactions occurred. Note that the use of new information aimed to give a better discrimination between fraud and non-fraud transactions. In accordance with Whitrow et al. (2009), a series of three suspicious transactions was likely to be an indicator of a fraud case.

We formed four data sets (denoted as D1, D2, D3, and D4) for evaluation, namely the original and three additional data sets. In line with the method in Bahnsen et al. (2016), D2 included additional aggregated features:

1. sum of transaction amounts for the same country and transaction type in the last 24 h.
2. sum of transaction amounts in the last 24 h;
3. no. of transactions having the same country and transaction type in the last 24 h;
4. no. of transactions in the last 24 h;

Four new features, in addition to the aggregated ones, were added to D3, as follows:

1. no. of transactions having the same type of device in the last 24 h;
2. no. of transactions having the same MCC (merchant category code) in the last 24 h;
3. sum of transaction amounts having the same type of device in the last 24 h.
4. sum of transactions amounts having the same MCC in the last 24 h;

D4 was produced using a GA (genetic algorithm) to determine the most significant features. Based on the survival-of-the-fittest concept, the GA is useful for undertaking search and optimization tasks. Here, feature selection with the GA was conducted using the default parameters in RapidMiner. During the feature selection process, the 'mutation' operator switched the features "on" and "off", while the 'crossover' operator interchanged the selected features. Further details on this process are given in RapidMiner (2018).

4.5 Experimental results

Table 9 lists the accuracy (ACC) rates of all experiments. In D1, the ACC rates were approximately 99%, with 11 out of 13 achieving more than 99.8%. NB produced the lowest accuracy rate of 32.8%. Improvement in ACC was achieved by using D2, where the accuracy rate of NB increased from 32.8 to 97.6%. The other 12 models achieved ACC rates of more than 99.9%. The best ACC rate was produced by DT and DS, which showed a minor improvement from that of D1. The D3 results from all models remained similar to those

Table 9 Accuracy (ACC) results (best in bold)

Methods	D1 (%)	D2 (%)	D3 (%)	D4 (%)
NB	32.878	97.615	97.472	97.971
DT	99.959	99.997	99.997	99.997
RF	99.959	99.996	99.993	99.996
GBT	99.847	99.990	99.990	99.988
DS	99.959	99.997	99.997	99.997
RT	99.959	99.968	99.969	99.985
ANN	99.959	99.990	99.988	99.990
MLP	99.959	99.966	99.969	99.975
LIR	99.959	99.959	99.959	99.959
LOR	99.958	99.972	99.971	99.975
SVM	99.959	99.965	99.966	99.969
RI	99.959	99.959	99.959	99.959
DL	95.113	99.984	99.985	99.993

from D2. The D4 results from all models showed either a minor increase or remained the same. DT and DS yielded the best ACC rate for all four data sets.

To further assess the performance, we employed the Matthews Correlation Coefficient (MCC) (Powers, 2011). MCC provides a performance indication in binary tasks. It yields a balanced metric in evaluating classification problems with different data sample sizes pertaining to the target classes. Both false and true negative and positive predictions are considered, as follows

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (13)$$

A total disagreement and a perfect prediction are indicated by -1 and $+1$, respectively.

The MCC scores are given in Table 10. In D1, the MCC scores were poor, with the best score of 0.12 achieved by GBT. Most of the remaining classifiers did not produce

Table 10 MCC rates (best in bold)

Methods	D1	D2	D3	D4
NB	0.013	0.119	0.115	0.129
DT	–	0.964	0.964	0.964
RF	–	0.945	0.909	0.946
GBT	0.120	0.882	0.882	0.869
DS	–	0.964	0.964	0.964
RT	–	0.648	0.591	0.835
ANN	–	0.882	0.863	0.882
MLP	–	0.704	0.732	0.767
LIR	–	–	–	–
LOR	–	0.749	0.729	0.756
SVM	–	0.709	0.717	0.732
RI	–	–	–	–
DL	0.032	0.794	0.821	0.906

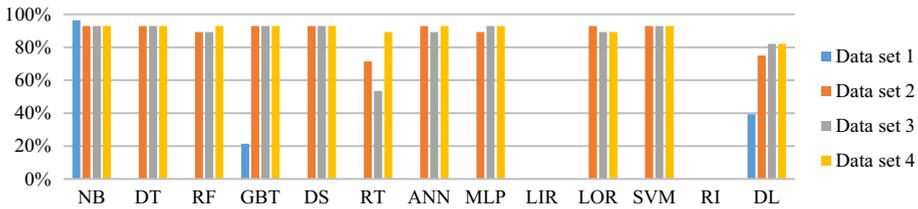


Fig. 1 Fraud detection rates of data sets 1–4

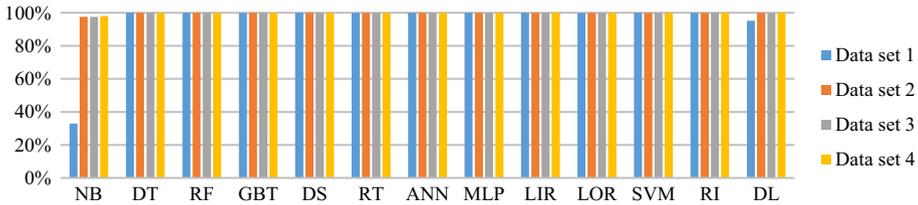


Fig. 2 Non-fraud detection rates of data set 1–4

desirable results, as their fraud detection rates were 0. A similar trend of MCC could be observed across D2, D3, and D4. DT and DS achieved the best MCC score of 0.964, while RF showed competitive MCC scores. Improvements in other classifiers could be observed as well, with differences in scores for D2, D3, and D4. The fraud detection rates are shown in Fig. 1. NB achieved the most stable performances across all four data sets. For other classifiers except NB, GBT, and DL, the fraud detection rate with D1 was zero. In D2, all classifiers, except LIR and RI, managed to detect fraud cases. This trend continued for both D3 and D4. It could be observed that feature aggregation was able to improve the fraud detection rates. The non-fraud detection rates are shown in Fig. 2. Most classifiers, except NB, produced perfect or almost perfect accuracy rates. These results were in contrast to the fraud detection rates shown in Fig. 1.

Table 11 summarizes the AUC scores. The results of D1 varied from 0.5 to 0.866, as achieved by GBT. In D2, improvements in most of the AUC scores could be observed. RF yielded the greatest improvement, i.e., from 0.5 to 0.958. All other classifiers achieved an increase in their AUC scores, except DT and RI. In D3, minor changes in the AUC scores could be observed. A similar observation could be made from the results of D4. While DS yielded the highest accuracy rates in Table 10, its AUC score was among the lowest, i.e., at 0.5 in all four data sets. One reason could be the simplicity of DS that contained a single split, which compromised its detection capabilities. GBT produced the highest AUC score of 0.967 for D2, D3, and D4. The boosting effect in GBT allowed learning to be focused on misclassified samples, which eventually led to a robust classification model, as shown in the results.

4.6 Managerial implications

From the perspective of financial institutions which provide payment cards services, the increase in fraud directly hits their business activities and impacts their profitability. With the rise in e-commerce transactions, the number of online transactions increases rapidly. The popularity of using payment cards is further fueled by the ability to perform online

Table 11 AUC results (best in bold)

Methods	D1	D2	D3	D4
NB	0.862	0.965	0.954	0.948
DT	0.500	0.500	0.500	0.500
RF	0.500	0.958	0.958	0.958
GBT	0.866	0.967	0.967	0.967
DS	0.499	0.500	0.500	0.500
RT	0.620	0.584	0.633	0.650
ANN	0.809	0.943	0.955	0.950
MLP	0.784	0.960	0.958	0.959
LIR	0.847	0.964	0.961	0.965
LOR	0.833	0.952	0.937	0.939
SVM	0.573	0.942	0.963	0.959
RI	0.500	0.500	0.500	0.500
DL	0.845	0.951	0.929	0.932

transactions from anywhere and anytime, with potentially lower prices in purchases. However, this comes with a cost, i.e., the number of fraud cases increases sharply as well. As a result, an effective fraud detection system for payment cards is vital for any financial institution to mitigate the risk of fraudulent transactions. It is costly for financial institutions to keep absorbing losses, as it creates financial uncertainty for them. The cost associated with efforts to handle risk and uncertainty will eventually be passed on to the retailers and consumers, leading to the increase in prices of goods. As an example, in a normal scenario, the payment card discount rate could be set at a lower range, closer to 1%. With an increase in the number of fraudulent transactions, the financial institution issuing the payment terminal may increase this rate to a higher rate, causing the merchant to get less from every single transaction. This leads to merchants selling goods online may increase the cost of the items, in which the consumers will end up paying for higher prices in their purchases.

In the recent year with the outbreak of Covid-19, consumers are having second thoughts on using cash. Using cash has been seen as potential hygiene issues, as the cash is passed around from one person to another. This has made many to move from handling cash to using plastic cards instead. While there are digital wallets or electronic wallets, which have gained popularity in recent times, their acceptance has made payment cards still a preferred choice for consumers. With the wide acceptance rate combined with consumers moving to cards from the Covid-19 pandemic, the rise of transactions will in turn have more fraud.

In this study, the developed fraud detection methods aim to mitigate the uncertainties caused by the issues discussed earlier. Indeed, it is crucial vital to detect fraudulent transactions before they happen, in order to ensure the overhead of business activities does not keep increasing. While there are many studies on payment card fraud detection, the use of real data is rare, since it is difficult to obtain real transaction records. As such, most of the existing methods use publicly available data sets for evaluation, and their effectiveness in real-world environments is unknown. In contrast, we have employed a real-world payment card database for demonstrating the usefulness of the developed methods with aggregated features. The resulting models can be readily used by financial institutions in real environment without requiring another round of assessment with the real data.

While there are many commercial fraud detection tools, fraudsters always try to outsmart the detection systems. As such, the journey of detecting and fighting fraud is a

continuous one, not just a one-off attempt. The developed methods offer a viable approach for further enhancement when more and more data samples are available for learning purposes, improving accuracy of the detection rates over time in real-world environments.

5 Conclusions

We have presented an investigation on fraud detection pertaining to payment card transactions in this paper. The main contributions are the development of a practical system utilizing aggregated features for payment cards fraud detection and the use of with real transaction records for evaluation and demonstration of the effectiveness of the developed system. Our study is important for reducing the risks of financial losses as well as uncertainties faced by institutions in their daily business activities. In our analysis, a total of thirteen statistical and machine learning methods, ranging from ANN to deep learning models have been used for evaluation. Three benchmark credit card data sets obtained from a public repository have been used for performance assessment. The AUC metric is employed, which indicates statistical difference in performance of various detection methods. The best AUC score achieved is 0.937 from GBT for the Australian data set.

Importantly, real-world payment card transactions, in addition to benchmark databases, have been employed in our study. The same statistical and machine learning methods have been used for performance assessment. Besides that, feature selection with the GA has been performed. Feature selection is key in ensuring that only important and non-redundant features are identified, to ensure the classification performance can be enhanced and, at the same time, computation load can be optimized for real-world implementation.

As it is onerous to acquire real financial records, the outcome of this study is significant in uncovering valuable insights into the robustness of machine learning algorithms in real-world environments. The features from the original data set are aggregated to form new features, with the aim to counter the effects of concept drift and enhance performance. Our findings pertaining to the benefits of feature aggregation methods are in line with some reported results in the literature, e.g. aggregation of data over a short period led to an increase in probability to detect fraud as indicated in Jha et al. (2012), while Lim et al. (2014) revealed that aggregation-based methods yielded better results as compared with those from standard transactions.

Based on the original features, the best AUC score (from GBT) is 0.866, while the best AUC score (also from GBT) increases to 0.967 with the use of aggregated features. RF has recorded the largest improvement in AUC, i.e., 0.5 with the original feature set to 0.958 with the aggregated features. In addition to AUC, another useful performance metric, i.e., MCC, has been adopted for evaluation. Again, the MCC scores indicate that aggregated features are able to improve the results. Both DT and DS produce the highest MCC score (0.964), GBT (which yields the highest AUC score) achieves 0.869. The results indicate the usefulness of the aggregated features in improving the overall performances in both ACC and MCC scores.

This study is significant in view of the rise in e-commerce activities whereby the number of online transactions increases rapidly in this digital era. Indeed, with the outbreak of Covid-19, consumers are now resorting to online purchases. As such, an effective fraud detection system for payment cards is vital for a financial institution to mitigate the risk of fraudulent transactions. The developed system, therefore, offers a viable solution for

financial institutions to detect fraudulent transactions pertaining to payment cards services in real environments.

In summary, we have evaluated the usefulness of machine learning and related models with aggregated features for fraud detection with both benchmark and real-world payment card database. The resulting models demonstrate a great potential for use by financial institutions in their daily business activities. Nevertheless, fraudsters always attempt to outsmart the detection systems. As such, the journey of detecting and fighting fraud is a continuous one. A number of limitations of this study and further research to enhance the develop fraud detection system are discussed.

6 Limitations and further research

The current study can be improved from several angles. Firstly, the real payment card database used is limited to a financial institution in Malaysia. The transactions are mostly occurred in the Asia region. It would be useful to acquire more real-world data from different regions, in order to fully evaluate the effectiveness of the developed method for detecting fraud in other regions around the world. The consumers in different regions may transact with different characteristics, and with varying spending patterns. As such, it is necessary to conduct further evaluation of the developed method with real data from different regions, in order to have a robust model that can be used for fraud detection of payment cards globally.

Another limitation is the use of single models for developing the fraud detection framework in this study. To further enhance the developed framework, hybrid models can be formed using combination of two or more models (Jiang et al., 2020). Hybrid models enable the use of more than one model to determine the transaction legitimacy, in order to improve further the fraud detection rate. In addition, online implementation of the detection methods will be investigated. This will allow detection and prevention of payment card fraud in real-time and in real environments. On the other hand, as financial risks differ between different regions, a variety of risk models and management strategies are available (Jawadi et al., 2019; Ben Amuer & Prigent, 2018), it is important to further improve the adaptability of the developed framework to suit various risk analysis methodologies. It is useful to investigate the applicability of different measurement errors in other financial domains (e.g. Ben Ameur et al., 2018, 2020), in order to ensure that the developed framework can be generalized for other financial risk analytic tasks.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

- Ameur, B., & Prigent, J. L. (2018). Risk management of time varying floors for dynamic portfolio insurance. *European Journal of Operational Research*, 269, 363–381.
- Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, 51, 134–142.

- Ben Ameer, H., Ftiti, Z., Jawadi, F., & Louhichi, W. (2020). Measuring extreme risk dependence between the oil and gas markets. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-020-03796-1>
- Ben Ameer, H., Jawadi, F., Idi Cheffou, A., & Louhichi, W. (2018). Measurement errors in stock markets. *Annals of Operations Research*, 262, 287–306.
- Bernard, P., De Freitas, N. E. M., & Mailet, B. B. (2019). A financial fraud detection indicator for investors: an IDEA. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-019-03360-6>
- Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602–613.
- Card Fraud Worldwide—The Nilson Report, vol. 1096 (2016).
- Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8), 3784–3797.
- Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10), 4915–4928.
- de Sá, A. G., Pereira, A. C., & Pappa, G. L. (2018). A customized classification algorithm for credit card fraud detection. *Engineering Applications of Artificial Intelligence*, 72, 21–29.
- Dua, D., & Graff, C. (2019). UCI machine learning repository [online]. Available <http://archive.ics.uci.edu/ml>. Irvine, CA: University of California, School of Information and Computer Science.
- Edge, M. E., & Sampaio, P. R. F. (2012). The design of FFML: A rule-based policy modelling language for proactive fraud management in financial data streams. *Expert Systems with Applications*, 39(11), 9966–9985.
- Everett, C. (2003). Credit card fraud funds terrorism. *Computer Fraud and Security*, 2003(5), 1.
- Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing*, 65, 139–151.
- Ferreira, F. A., & Meidutė-Kavaliauskienė, I. (2019). Toward a sustainable supply chain for social credit: Learning by experience using single-valued neutrosophic sets and fuzzy cognitive maps. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-019-03194-2>
- Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479, 448–455.
- Forbes. (2011). Bringing trust back to the table—part one: Adyen and Mobile Payments.
- Forough, J., & Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. *Applied Soft Computing*, 99, 106883.
- Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In *International conference on neural information processing* (pp. 483–490). Cham: Springer.
- Gómez, J. A., Arévalo, J., Paredes, R., & Nin, J. (2018). End-to-end neural network architecture for fraud scoring in card payments. *Pattern Recognition Letters*, 105, 175–181.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.
- Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing*, 69, 541–553.
- Jawadi, F., Louhichi, W., Cheffou, A. I., & Ben Ameer, H. (2019). Modeling time-varying beta in a sustainable stock market with a three-regime threshold GARCH model. *Annals of Operations Research*, 281, 275–295.
- Jha, S., Guillen, M., & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert Systems with Applications*, 39(16), 12650–12657.
- Jiang, C., Song, J., Liu, G., Zheng, L., & Luan, W. (2018). Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism. *IEEE Internet of Things Journal*, 5(5), 3637–3647.
- Jiang, M., Jia, L., Chen, Z., & Chen, W. (2020). The two-stage machine learning ensemble models for stock price prediction by combining mode decomposition, extreme learning machine and improved harmony search algorithm. *Annals of Operations Research*, 1, 1–33. <https://doi.org/10.1007/s10479-020-03690-w>
- Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P. E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100, 234–245.

- Li, Z., Huang, M., Liu, G., & Jiang, C. (2021). A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection. *Expert Systems with Applications*, *175*, 114750.
- Lim, W. Y., Sachan, A., & Thing, V. (2014). Conditional weighted transaction aggregation for credit card fraud detection. In *IFIP International conference on digital forensics* (pp. 3–16). Berlin: Springer.
- Lucas, Y., Portier, P. E., Laporte, L., He-Guelton, L., Caelen, O., Granitzer, M., & Calabretto, S. (2020). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. *Future Generation Computer Systems*, *102*, 393–402.
- Pavía, J. M., Veres-Ferrer, E. J., & Foix-Escura, G. (2012). Credit card incidents and control systems. *International Journal of Information Management*, *32*(6), 501–503.
- Powers, D. (2011). Evaluation: From prediction, recall and F-factor to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.
- Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. *IEEE Access*, *6*, 14277–14284.
- RapidMiner. (2018). Optimize Selection (RapidMiner Studio Core) [Online]. Available https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection.html.
- Robinson, W. N., & Aria, A. (2018). Sequential fraud detection for prepaid cards using hidden Markov model divergence. *Expert Systems with Applications*, *91*, 235–251.
- Rtayli, N., & Enneya, N. (2020). Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization. *Journal of Information Security and Applications*, *55*, 102596.
- Russac, Y., Caelen, O., & He-Guelton, L. (2018). Embeddings of categorical variables for sequential data in fraud context. In *International conference on advanced machine learning technologies and applications* (pp. 542–552). Cham: Springer.
- Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916–5923.
- Sariannidis, N., Papadakis, S., Garefalakis, A., Lemonakis, C., & Kyriaki-Argyro, T. (2020). Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: Decision making based on machine learning (ML) techniques. *Annals of Operations Research*, *294*(1), 715–739.
- Seera, M., Lim, C. P., Tan, S. C., & Loo, C. K. (2015). A hybrid FAM–CART model and its application to medical data classification. *Neural Computing and Applications*, *26*(8), 1799–1811.
- Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38–48.
- Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, *18*(1), 30–55.
- Zhang, J., Xiong, Y., & Min, S. (2019). A new hybrid filter/wrapper algorithm for feature selection in classification. *Analytica Chimica Acta*, *1080*, 43–54.
- Zhang, X., Han, Y., Xu, W., & Wang, Q. (2020). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information Sciences*. <https://doi.org/10.1016/j.ins.2019.05.023>
- Zhu, H., Liu, G., Zhou, M., Xie, Y., Abusorrah, A., & Kang, Q. (2020). Optimizing Weighted Extreme Learning Machines for imbalanced classification and application to credit card fraud detection. *Neuro-computing*, *407*, 50–62.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Manjeevan Seera¹  · Chee Peng Lim² · Ajay Kumar³ · Lalitha Dhamotharan⁴ · Kim Hua Tan⁵

Chee Peng Lim
chee.lim@deakin.edu.au

Ajay Kumar
akumar@em-lyon.com

Lalitha Dhamotharan
L.Dhamotharan@exeter.ac.uk

Kim Hua Tan
Kim.Tan@nottingham.ac.uk

¹ Econometrics and Business Statistics, School of Business, Monash University Malaysia, Selangor, Malaysia

² Institute for Intelligent Systems Research and Innovation, Deakin University, Geelong, VIC, Australia

³ AIM Research Center on Artificial Intelligence in Value Creation, EMLYON Business School, Écully, France

⁴ University of Exeter Business School, University of Exeter, Exeter, UK

⁵ Nottingham University Business School, Nottingham, UK