Applied Mathematics and Computational Intelligence
Volume 11, No.1, Dec 2022 [88-103]

Fraud Detection by Machine Learning Techniques

Esraa Faisal Malik[1], Khai Wah Khaw[2*], XinYing Chew[3]

[1]School of Management, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
[2] School of Management, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
[3] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia

[*] Corresponding author: khaiwah@usm.my

Received: 1 December 2021; Accepted: 28 February 2022; Available online (In press): 7 March 2022

ABSTRACT

Over the years, the negative impact of financial fraud on organizations and countries have been increasing significantly. Conventional methods such as expert's judgment are usually used to detect financial fraud. However, these methods suffer from serious drawbacks due to time consumption, human errors and high operational cost. Hence, the need for automating the fraud detection method arises. Researchers have been using machine learning to detect fraudulent cases, this approach has been widely used in financial fraud detection to address the shortcomings of conventional methods as it automates the detection process and has the potential to resolve the disparity between the impact of fraud detection and its efficacy. This research applies two feature selection methods (correlation and wrapper) with three different machine learning algorithms (Naïve Bayes, Support Vector Machine and Random Forest) to determine the optimum algorithm that can efficiently classify fraudulent and non-fraudulent firms based on a real-life dataset from the Auditor General Office of India between 2015 and 2016. A data science life cycle in machine learning was adopted. The results and evaluation revealed that machine learning algorithms had a superior advantage and can be used over the traditional methods of fraud detection.

Keywords: Algorithms; fraud detection; feature selection; machine learning

1 INTRODUCTION

Over the past few years, great focus has been put into the evolution of financial technologies such as cryptocurrency, crowdfunding platforms and mobile payments in order to help people and financial institutions in their daily/business transactions. However, the major obstacle of these technologies is that it has facilitated financial fraud. Financial Fraud is the deliberate use of illegal techniques and activities to achieve financial gains [1]. In previous studies, it was found that there are several types of financial fraud such as; insurance fraud [2], securities and commodities fraud, credit card fraud [3]–[5], financial statement fraud [1], [6]–[8] and money laundering [9], [10]. Financial fraud costs developed countries billions of dollars annually. In 2007, BBC News reported that the UK lost around 1.6 billion pounds due to fraudulent insurance [11]. Others estimated that fraud costs the US industry more than 400 billion dollars yearly [12], [13]. Recently, financial statement fraud was discovered in large companies like Satyam, WorldCom, Enron and Lucent [2]. Although there has been a long-standing interest in fraud detection, yet it starts to become an international problem affecting both

global and domestic markets. According to [14] this issue has risen to the second-highest level in the world over the last 20 years, with 47 percent of businesses experiencing fraud offences and damages totaling 42 billion dollars in the last two years. Additionally, the authors in [15] indicated that customer fraud (e.g., identity theft, credit card fraud, mortgage fraud, etc.) has the highest occurrence rate for fraud incidents, which equaled 35 percent while tax fraud had less than 10 percent. The problem of fraud caught the governments and financial institutions concerns, not only because of the monetary losses but because these acts can seriously harm organizations reputation as they are responsible for the sudden failure of many reputable institutions [15]. Moreover, according to the Association of Certified Fraud Examiners (ACFE), fraud events can occur across a wide range of private and public institutions as well as throughout the economy. Furthermore, government, public administration, financial services and banking, and manufacturing, were particularly vulnerable to the fraudulent indictment, however, they pointed out that the high fraud rate in these areas does not necessarily imply that there is more fraud in these industries; rather, it could simply indicate that companies in these industries employ more CFEs (Certified Fraud Examiners) than others [16]. As a result, the need to uncover and report financial fraud had increased. Previously, the organizations used to depend on financial experts' judgments and knowledge to detect financial fraud cases. However, this traditional method has several drawbacks such as time consumption, presence of human errors and high operational cost [17]. Therefore, a new term called financial fraud detection (FFD) received great attention. FFD is a process of identifying fraudulent financial data using realistic data by uncovering hidden patterns of fraudulent activities using machine learning techniques [11].

Machine learning is the science of getting computers to learn without being explicitly programmed [18]. This method has been widely applied in financial fraud detection to overcome the limitations of the traditional methods as it automates the detection process and has the capability to solve the inconsistency between the impact of fraud detection and its effectiveness [19]. There are three major types of machine learning; supervised learning, unsupervised learning, and semi-supervised learning where each one is supported by different algorithms, used for specific tasks, and can be employed in a certain dataset. For instance, in the credit card fraud detection domain, the supervised learning, the dataset is labelled into legitimate and fraudulent transactions. While in unsupervised learning, the cardholder's past transactions are used to model the spending behaviour of the cardholder. Where a coming transaction is considered possible fraudulent when it does not match the existing behaviour model. On the other hand, semi-supervised learning is a combination of the two supervised learning and unsupervised learning. As mentioned in the review papers [20], [21] the classification machine learning approach has been the most popluar method to detect fraudelnt activites. The new predicted labels are unordered, predefined, and discrete. Hence, machine learning classification algorithms will be used. Classification is the process of recognizing a set of mutual features that differentiate data classes and concepts [19]. Additionally, Previous researcher confirms that machine learning classifiers has been widely used in different domains, especially in fraud detection such as Support Vector Machine (SVM)[7], [22]–[27], Logistic Regression (LR) [7], [23], [28], Decision Tree (DT) [7], [29], Naïve Bayes (NB) [30]–[33] K-nearest Neighbor (KNN) [5], [28], [34]. Neural Network (NN) [23], [35] and Random Forest (RF) [36]–[40]. Nevertheless, the three most common classifer in the literature were SVM, NB and RF as shown in Table 1 [20]. As well, each one of these classifier have an advantges (table 2)

Table 1: Classification of data mining techniques based on their fraud types [20]

No	Method	Frequency	Fraud types	
			Bank fraud	
			Credit card fraud	Internet bank fraud
1	SVM	17	9	
2	Decision Tree	6	3	
3	Logistic Regression	9	6	
4	Outliers Detection	7	1	
5	Bayesian Network	2	1	
6	K-Nearest Neighbor	8	4	
7	Neural Network	10	6	
8	Genetic Algorithm	5	4	
9	Hidden Markov Model	6	3	3
10	Fuzzy Logic	4	2	1
11	Danger Theory	1	1	
12	Naïve Bayes	11	5	
13	GMDH	2	1	
14	Gradient Boosting Tree	4	3	
15	Random Forest	11	4	

Table 2: . Comparison between RF, NB and SVM strength

Algorithm	Strength
RF	Ease of use, interpretability and provide a strong indicator of which features are most relevant for the classification model
NB	Robust to noisy and irrelevant data points, computationally efficient and easily understood
SVM	High tolerance to noisy features, effective in high dimensional spaces and memory-efficient

As a result of their strength and popluartly in the literatures, therefore, in this research a comparative study between SVM, NB and RF machine learning classifications techniques was made to assist the auditors in selecting the most appropriate classifier for detecting fraudulent companies. As this is the aim of this research, to aid financial institutions improving fraud detection utilizing the use of several machine learning models -SVM, NB, and RF- on the company's automated detection system. Therefore, by leveraging this, financial institutions and banks will be able to undertake their business to a higher level as they will cut computational cost, and lower false alarm as well as fraud rate.

The remainder of this article is organized as follows; Section II describes the methodological framework and experimental setup for the research. Section III presents the results and discussion. Section IV concludes the research and recommend further future studies.

2 METHODOLOGICAL FRAMEWORK AND EXPERIMENTAL SETUP

A conceptual framework for machine learning using data science life cycle (DSLC) was implemented in this study [41]. DSLC is not exclusively for a specific application or an industry which makes it the predominantly used method for data mining projects across all industries as it assists in the project's success and yield beneficial results [42]. It revolves around machine learning utlization and other analytical approacges to derive insights and predictions from data in order to achieve the intended outcome and assist decision-makers in their final decisions. The DSLC consists of five different phases as shown in Fig. 1. In the first phase, we asked two questions:

1. to what extent can machine learning techniques predict fraud detection?
2. what are the best algorithms to perform this task?

In the second phase, data were collected from the Indian audit dataset of firms that is publicly available in the UCI machine repository. The dataset was obtained from the Auditor General Office (AGO) of India between 2015 and 2016 [43]. There is a total of 776 instances and 26 attributes which are all numeric. The used dataset contains firms from 14 different sectors. Table 3 lists the sectors and the counts of firms.

In the third phase, an initial exploratory analysis of the dataset was conducted in Weka software. We removed three different attributes (location ID, total risk, and detection risk) as they were irrelevant and had no contribution to the predictive model. Then, the exploratory analysis displays one missing value in the numeric attribute Money_value; therefore, we replaced it by the mean using the filter ReplaceMissingValues [44]. Moreover, it was shown the target predicted class “Risk” as a numeric attribute represented by 0 and 1. Since the purpose of this work is to classify fraudulent and non-fraudulent firms; the data type was converted from “numeric” to “nominal” which is more appropriate for the subsequent processes and the use of machine learning classification algorithms. Furthermore, as illustrated in Fig. 2 the class distribution of Risk is 471 (61%) and 305 (39%) for non-fraudulent (represented by 0) and Fraudulent (represented by 1). There is a noticeable imbalanced problem in the used dataset since the identified fraudulent firms are significantly less than non-fraudulent ones. However, the issue of the imbalance in the used dataset is not severe compared to some previous studies [45], so it is better to use it as available without any sampling. After the data preprocessing stage, the dataset used to build the predictive model now consists of 23 attributes and 776 instances. In the fourth phase before proceeding with the modelling using the three different techniques; NB, SVM and RF. an experimental setup and feature selection dimensionality reduction were conducted.

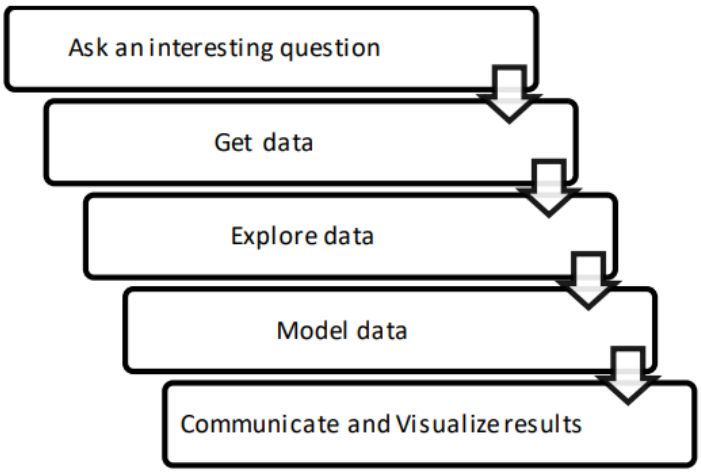

Figure 1: Data science life cycle

Table 3: The sectors listed in the dataset and the count of firms

Sector	Number of firms
Agriculture	200
Fisheries	41
Industries	37
Tourism	1
Land	5
Communication	1
Science and Technology	3
Electrical	4
Animal Husbandry	95
Public Health	77
Forest	70
Buildings and Roads	82
Corporate	47
Irrigation	114

2.1 Experimental Setup

2.1.1 Performance Metric Selection

Traditionally, accuracy measure is used to determine the performance of the predicted model, but because of the imbalanced data used in this research, this measure will not be efficient. Cases predicted as possible fraud are cost in fraud detection, as they are token up for further investigation. Precise detection of cases of fraud helps to avoid costs resulting from fraudulent activity, which is usually greater from fraudulent activity than the cost of investigation of possible fraudulent activity. Therefore, a Type-I error and Type-II error will be used. Type-I error (false positive) provides the total of nonfraudulent firms that are mistakenly labelled as fraudulent. Whereas Type-II error (true negative) indicates the sum of nonfraudulent firms that are incorrectly labelled as fraudulent [46]. In addition, recall and precision were also suitable to evaluate the predictive model whether it is capable to identify fraudulent firms accurately. A recall is the proportion of real fraudulent firms predicted correctly by the model as fraudulent. On the other hand, precision is the proportion of

predicted observations as fraudulent firms by the model that is a real fraudulent [47]. If the recall is approximately equal to 1 this means that all the firms are classified as fraudulent. On the contrary, precision will be low because many non-fraudulent firms are classified wrongly. Consequently, performance measure like F-measure gives equal consideration to precision and recall [48] Refer to "(1)"

$$F - measure = 2 * ((precision * recall)/(precision + recall)) \quad (1)$$

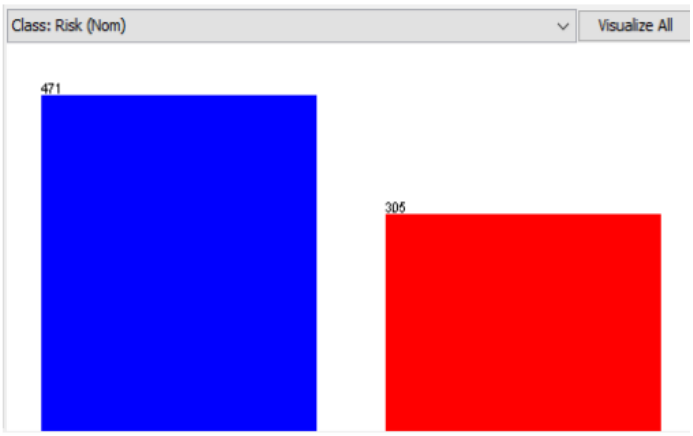

Figure 2: Class distribution of the attribute Risk

2.1.2 Test Option Selection

The dataset of 776 instances is sufficient to develop a classification model. The cross-validation option was selected since the used dataset is relatively small to train and test split. K fold cross-validation method (K=10) is applied to check the robustness of the building model [46].

2.2 Dimensionality Reduction

As the dataset consisted of many attributes; dimensionality reduction was performed to avoid computation cost for model development. According to [49] dimensionality reduction is the process of reducing the dimension of the dataset into meaningful representations. The dimensionality reduction techniques are divided into two, feature selection and feature extraction. Feature selection is a process of identifying and selecting relevant features to a sample of the dataset. Whereas feature extraction is a process of generating new features from already existing ones. In this research, we performed feature selection using Weka. Feature selection is already a function in Weka that is divided into two parts: attribute evaluator and search method. Some of the attribute evaluator techniques require the use of specific search methods such as WrapperSubsetEval (learner-based feature selection technique) and CorrelationAttributeEval. Both techniques require the use of GreedyStepwise, BestFirst or Ranker search methods. In this work, the search method used in the Wrapper technique is BestFirst. Whereas in the CorrelationAttributeEval technique, the Ranker

search method was selected. Performance evaluation of the algorithms on the dataset with different subsets of selected attributes was conducted. Before selecting the features from the 23 attributes, we generated a preliminary performance using all attributes and NB as the machine learning algorithm baseline. The results showed 0.972, 0.898 and 0.934 for Fraudulent class in precision, recall and F-measure, respectively. The results of the correlation of each attribute with a class label using correlation feature selection methods are shown in Table 4. An obvious gap between 0.62 (Score_A) and 0.416 (CONTROL_RISK) was noticed. Hence, 0.5 was set as a cut off value and therefore four attributes with correlation > 0.5 were selected as the first set of feature selection. The second set of feature selection was selected using the learner wrapper feature selection method. To identify which attribute is highly correlated with the class label Risk, three different classifiers; NB, SVM and RF were used. As illustrated in Table 5, three sets of features were used in the subsequent experiments. Next, a predictive model was developed using the mentioned classifiers.

Table 4: Correlation ranked attributes

Ranking	Attribute
0.786	**Score**
0.688	**Score_MW**
0.636	**Score_B**
0.620	**Score_A**
0.416	CONTROL_RISK
0.412	Risk_E
0.404	District_LOSS
0.394	Sector_score
0.385	Risk_A
0.379	PARA_A
0.357	Inherent_Risk
0.342	Risk_C
0.308	numbers
0.299	Prob
0.257	PARA_B
0.257	Money_Value
0.255	Risk_B
0.254	Risk_D
0.239	History
0.217	Audit Risk
0.215	Risk_F
0.177	PROB

Table 5: Wrapper selected attributes

NB	SVM	RF
Score_A	Sector_score	Audit Risk
District_Loss	PARA_A	
Risk_E	Score_A	
Inherent_Risk	Risk_A	
CONTROL_RISK	Risk_C	
Audit_Risk	Score_MW	
	Risk_E	
	Prob	
	Score	
	CONTROL_RISK	

2.3 Machine Learning Algorithms

2.3.1 Naïve Bayes

It is considered as one of the simplest statistical algorithms in the Bayesian classifiers in which it performs probabilistic prediction. This algorithm is based on the Bayes Theorem that describes the probability of an event based on prior knowledge of conditions that might be related to the event [50], Refer to "(2)". Rather than being a single distant algorithm, it is a set of algorithms that work on the underlying principle "The value of a given feature is independent of the value of any other feature". It allows the class conditional independence, in which it assumes that all the features in the data are independent given class [51] and each class label and attribute are random variables. NB set the class to the new observation by calculating the highest attributes for each class given the values of the attributes [51].

$$P(A|B) = (P(B|A)\ P(A))/\ P(B) \quad (2)$$

- P (A|B): is the posterior probability; the probability of class A given the attribute(s) B.
- P(B|A): is the LIKELIHOOD; probability of attribute(s) B given that class A was true.
- P(A): is the probability of class A being true (regardless of the attribute(s)). This is called the prior probability of A.
- P(B) Is the Prior probability of predictor.

NB is considered an effective machine learning classification classifier that is highly computationally efficient, easily understood and has the ability to interpret results [6]. In addition, it requires less data for the training -as in the case of this study-, it's practically suitable when the dimensionality of inputs is high and it's robust to irrelevant and noisy attributes [2]. When comparing conventional classifiers such as LR, DT, KNN to NB, the results indicate that the remarkable results can be accomplished by Naïve Bayes [30]. Furthermore, it has performed effectively in the financial fraud detection field [4], [6].

2.3.2 Support Vector Machine

SVM is another algorithm that is used for supervised learning. It supports both regression and classification modelling and can handle various numbers and types of attributes. SVM is based on the Structural Risk Minimization (SRM) principle from statistical learning theory [5]. As shown in Fig.3; there are three components of SVM representation; hyperplane, support vector and support vector machine. A hyperplane is a line that splits the input variable space by its class in a multidimensional space. On the other hand, the support vector is the nearest data point to the hyperplane. In addition, the support vector machine is a hyperplane that efficiently separates the two classes. In order to find the optimum hyperplane, two conditions must be applied: (1) to choose a hyperplane that optimally segregates the classes. (2) to maximize the distance from the closest support vector [50]. There are two types of SVM: linear SVM and non-linear SVM. Each type will perform the modelling for non-linear and linear data in a high dimensional space, respectively. For the non-linear SMV there is various type of kernels that can be used such as linear, polynomial, radial basis function (RBF), and sigmoid. RBF kernel is the most popularly used and highly recommended since it has unlimited response across the entire range of the real x-axis.

The support vector machine has been widely used in different areas. In [35] they used SVM to extract both sentiment and concept knowledge from a corpus to investigate the performance of document

classification for a set of annual reports and it has the best accuracy. Whereas in [52] it was used for detecting money laundering along with other algorithms. Additionally in 2009, a study was performed using SVM, RF LR, NB and CART to develop a transaction aggregation as a strategy for credit card fraud detection [4]. Similar to Naïve Bayes, SVM has a high tolerance for noisy data. Moreover, as a result of the hyperplane function; SVM shows an effective performance and high generalization capability in the machine learning classification problems [39].

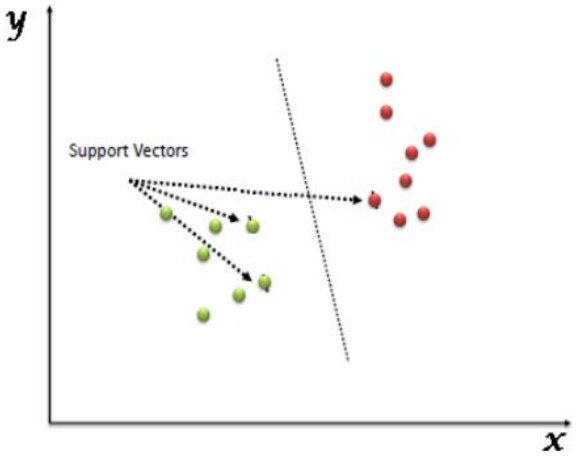

Figure 3: SVM classifier [37]

2.3.3 Random Forest

RF is an ensemble of regression trees or classification [18] as it boosts the decision tree by using the multiple decision tree voting mechanisms. For illustration, if there are N samples and M variables), the RF classifier mechanism m will be to first identify the m value that is used to determine the number of chosen variables by each tree classifier. The second is to generate k tree classifiers by obtaining and using k samples from the dataset and for testing purposes K bags of external data are created. Third, each tree classifier will classify it after entering the classified sample and then the classification result will be determined by all classifiers as per the majority rule [7]. Ensemble methods usually perform well when individual members are different and random. The random forest can acquire differences between individual trees using two sources for randomness: (1) in the training data, each tree is constructed on a different bootstrapped sample. (2) When constructing the individual tree, only a data attributes subset that is selected randomly is chosen at each node. RF incorporates the principles of bagging in which individual models are an ensemble by sampling with replacement from the training data, and the random subspace approach in which each tree in an ensemble is constructed from a random subset of attributes [39]. RF has been popular in recent years. They are easy to use, require only two adjustable parameters: the number of trees (T) in the ensemble and the attribute subset size (b) [18]. Several studies have proved that random forest shows an overall high performance when compared to other algorithms using different datasets in fraud detection [3], [7].

3 RESULTS AND DISCUSSION

In the last phase, the possibility of obtaining a higher F-measure or obtaining the same performance with fewer attributes in order to reduce computation cost was explored. For this purpose, experiments with three different machine learning algorithms were performed with different parameters and three feature selection sets. One parametric algorithm (NB classifier) and two non-parametric algorithms (SVM and RF) in the experiments were selected. As discussed earlier, we proposed the NB algorithm with default parameters as our baseline model in the exploratory stage. The result was 0.972, 0.898 and 0.934 for the Fraudulent class in precision, recall and F-measure, respectively. The result indicates that it is a good baseline to serve as a base for a class label in precision, recall and F-measure for comparison in a subsequent experiment. The summary of the results of the experiment is shown in Table 6.

3.1 Experiment 1: Naïve Bayes

One of the parametric algorithms, Naïve Bayes applied in this experiment with tuning parameters; useKernelEstimator and useSupervisedDiscretization. The results indicated that using the useSupervisedDiscretization parameter on the Wrapper feature selection set increased the efficiency of the NB model with an F-measure equal to 1.00. Furthermore, Type-I and Type-II error was found to be equal to zero which is significantly better than the results from the correlation feature selection set. The low performance of NB in the correlation set can be a cause of the independent assumptions; meaning that the attributes are dependent on each other.

3.2 Experiment 2: Support Vector Machine

The performance of SVM indicates a relatively low performance with an F-measure of 0.852 and 0.845 for correlation and wrapper, respectively. On the other hand, the default dataset obtained an F-measure result of 0.955. Hence, the model with the original dataset and default settings is considered as the optimum model in Experiment 2 with the closest F-measure value to 1.00.

3.3 Experiment 3: Random Forest

The performance of the tree classifier RF was further explored in this experiment. No parameter tuning was performed on the three different sets. The best RF classifier model in this experiment was achieved when using the wrapper feature selection set as attributes to build a classification model. The F-measure for class label Fraudulent was found to be exactly 1.00.

As clearly evident in Table 4, based on the findings from the experiments, it was found that this problem is more favourable to NB and RF. The most effective feature selection set is a wrapper with 6, 1 attributes in the NB and RF models, respectively. Whereas, for SVM feature selection methods had a negative effect on the algorithm performance efficiency as it lowered the F-measures using both Correlation and Wrapper techniques with values of 0.852 and 0.845, respectively. The best model for each experiment is illustrated in Table 4. Generally, the chosen machine learning algorithms performed better than the baseline algorithm using default parameters and the original dataset. Furthermore, two models using NB and RF applying the wrapper technique were selected as the best predictive model in this classification problem as they produce the highest F-measures. Nevertheless, RF has the advantage of obtaining more descriptive information and the ability to provide more illustrative information on attributes that differentiate fraudulent and non-fraudulent

firms. Potentially, at a more critical level, information obtained from the tree classifier is influential for the Auditors who aim to understand their customers' behaviour and how to indicate fraudulent activities.

Table 6: Summary of results

Algorithm	Attribute [1]	Parameter [2]	Precision [3]	Recall [3]	F-measure [3]	Type –I error	Type – II error
Naïve Bayes	All	Default	0.972	0.898	0.934	2.837 %	6.275 %
		useKernelEstimator=True	1.000	0.872	0.932	0.000 %	7.647 %
		useSupervisedDiscretization=True	0.993	0.993	0.993	0.656 %	0.425 %
	Correlation	Default	0.931	0.836	0.881	6.934 %	9.960 %
		useKernelEstimator=True	0.914	0.836	0.873	8.602 %	10.060%
		useSupervisedDiscretization=True	0.834	0.875	0.854	16.562%	8.333 %
	Wrapper	Default	0.997	0.987	0.992	0.331 %	0.844 %
		useKernelEstimator=True	1.000	0.961	0.980	0.000 %	2.484 %
		useSupervisedDiscretization=True	**1.000**	**1.000**	**1.000**	**0.000 %**	**0.000 %**
SVM	**All**	**Default**	**0.976**	**0.934**	**0.955**	**2.397 %**	**9.827 %**
	Correlation	Kernel=RBFKernel filterType= Normalize training data	0.963	0.764	0.852	3.719 %	13.483 %
	Wrapper	Kernel=RBFKernel filterType= Normalize training data	0.983	0.741	0.845	1.739 %	14.469 %
Random forest	All	Default	1.000	0.997	0.998	0.000 %	0.212 %
	Correlation	Default	0.969	0.931	0.950	3.072 %	4.348 %
	Wrapper	**Default**	**1.000**	**1.000**	**1.000**	**0.000 %**	**0.000 %**

[1] Correlation is the selected 4 attributes from Correlation Feature Selection output. "Wrapper" is the selected 6, 10, 1 attribute from Leaner Feature Selection output for NB, SVM and RF, respectively. [2] Parameter that produces highest F-measure during experiments. [3] for the class label of (1) only.

4 CONCLUSION

Machine learning is the science of getting machines to learn without providing them with clear instructions [18]. Recently, it has been used in different areas in the financial sector. In this study, an effort has been made to help the auditors to find the most appropriate algorithm for the fraudulent detection problem by following a DSLF consisting of; ask an interest question, obtaining data, exploring data, model, and interpreting results. This paper inspected the performance of three machine learning techniques: NB, SVM and RF for fraud detecting cases. A real-life dataset from AGO of India between 2015 and 2016 was used for model building and evaluation. The results indicated that NB and RF have achieved the best performance equally when compared to SVM. Nevertheless, the model created by RF was selected as the best model as it can provide descriptive information. This study has one major limitation: the relatively small size of the dataset. Consequently, for future studies, a large dataset must be used. On the other hand, a hybrid model can be developed to compare with the results obtained from this study.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Z. Wei and K. Gaurav, "Detecting evolutionary financial statement fraud," *Decis. Support Syst.*, vol. 50, no. 3, pp. 570-575, 2011, doi: 10.1016/j.dss.2010.08.007.

[2] A. Sharma and P. Kumar Panigrahi, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," *Int. J. Comput. Appl.*, vol. 39, no. 1, pp. 37–47, 2012, doi: 10.5120/4787-7016.

[3] N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," *Decis. Support Syst.*, vol. 95, no. January, pp. 91–101, 2017, doi: 10.1016/j.dss.2017.01.002.

[4] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data Min. Knowl. Discov.*, vol. 18, no. 1, pp. 30–55, 2009, doi: 10.1007/s10618-008-0116-z.

[5] O. Ata and L. Hazim, "Comparative analysis of different distributions dataset by using data mining techniques on credit card fraud detection," *Teh. Vjesn.*, vol. 27, no. 2, pp. 618–626, 2020, doi: 10.17559/TV-20180427091048.

[6] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods," *Knowledge-Based Syst.*, vol. 128, pp. 139–152, 2017, doi: 10.1016/j.knosys.2017.05.001.

[7] J. Yao, Z. Jie, and L. Wang, "A Financial Statement Fraud Detection Model Based on Hybrid Data Mining Methods," in *2018 Int. Conf. Artif. Intell. Big Data*, pp. 57–61, 2018, doi: 10.1109/ICAIBD.2018.8396167.

[8] C. C. Lin, A. A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," *Knowledge-Based Syst.*, vol. 89, pp. 459–470, Nov. 2015, doi: 10.1016/j.knosys.2015.08.011.

[9] M. Alkhalili, M. H. Qutqut, and F. Almasalha, "Investigation of Applying Machine Learning for Watch-List Filtering in Anti-Money Laundering," *IEEE Access*, vol. 9, pp. 18481–18496, 2021, doi: 10.1109/ACCESS.2021.3052313.

[10] M. Jullum, A. Loland, R. B. Huseby, G. Anonsen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," *J. Money Laund. Control*, vol. 23, no. 1, pp. 173–186, Jan. 2020, doi: 10.1108/JMLC-07-2019-0055.

[11] E.W.T. Ngai, Y. Hu, Y.H. Wong, Y. Chen and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559-569, 2011. doi: 10.1016/j.dss.2010.08.006.

[12] J. T. Wells, "Occupational fraud and abuse," in *Corporate Fraud Handbook: Prevention and Detection*, Fifth Edit., 2017, pp. 366– 379.

[13] C. Spathis, M. Doumpos, and C. Zopounidis, "Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques," *Eur. Account. Rev.*, vol. 11, no. 3, pp. 509–535, Sep. 2002, doi: 10.1080/0963818022000000966.

[14] P. G. E. C. and F. Survey, "Fighting fraud : A never-ending battle," 2020.

[15] S. Sule, S. S. Ibrahim, and A. A. Sani, "The Effect of Forensic Accounting Investigation in Detecting Financial Fraud : A Study in Nigeria," *International Journal of Academic Research in Business and Social Sciences*, vol. 9, no. 2, pp. 545-553, 2019, doi: 10.6007/IJARBSS/v9-i2/5590.

[16] Association of Certified Fraud Examiners (ACFE), "Report to the nations on occupational fraud and abuse: 2020 global fraud study," *Acfe*, p. 88, 2020.

[17] J. West and M. Bhattacharya, "Intelligent Financial Fraud Detection : A Comprehensive Review," *Comput. Secur.*, 2015, doi: 10.1016/j.cose.2015.09.005.

[18] S. Vieira, W. H. Lopez Pinaya, and A. Mechelli, *Introduction to machine learning*. The MIT Press, 2019.

[19] S. Wang, "A comprehensive survey of data mining-based accounting-fraud detection research," in *2010 Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2010*, vol. 1, pp. 50–53, 2010, doi: 10.1109/ICICTA.2010.831.

[20] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Comput. Sci. Rev.*, vol. 40, p. 100402, 2021, doi: 10.1016/j.cosrev.2021.100402.

[21] M. N. Ashtiani and B. Raahemi, "Intelligent Fraud Detection in Financial Statements using Machine Learning and Data Mining: A Systematic Literature Review," *IEEE Access*. 2021, doi: 10.1109/ACCESS.2021.3096799.

[22] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLoS One*, vol. 12, no. 7, p. e0180944, Jul. 2017, doi: 10.1371/JOURNAL.PONE.0180944.

[23] S. Ben Jabeur, A. Sadaaoui, A. Sghaier, and R. Aloui, "Machine learning models and cost-sensitive decision trees for bond rating prediction," *J. Oper. Res. Soc.*, vol. 71, no. 8, pp. 1161–1179, 2020, doi: 10.1080/01605682.2019.1581405.

[24] R. D. Camino, R. State, L. Montero, and P. Valtchev, "Finding suspicious activities in financial transactions and distributed ledgers," in *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 2017-Novem, pp. 787–796, Dec. 2017, doi: 10.1109/ICDMW.2017.109.

[25] N. V Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," *Data Min. Knowl. Discov.*, pp. 225–252, 2008, doi: 10.1007/s10618-008-0087-0.

[26] M. Kirlidog and C. Asuk, "A fraud detection approach with data mining in health insurance," *Procedia-Social Behav. Sci.*, vol. 62, pp. 989–994, 2012, doi: 10.1016/j.sbspro.2012.09.168.

[27] X. Li and S. Ying, "Lib-SVMs detection model of regulating-profits financial statement fraud using data of chinese listed companies," in *2010 Int. Conf. E-Product E-Service E-Entertainment, ICEEE2010*, 2010, doi: 10.1109/ICEEE.2010.5660371.

[28] A. Phongmekin and P. Jarumaneeroj, "Classification models for stock's performance prediction: A case study of finance sector in the stock exchange of Thailand," Aug. 2018, doi: 10.1109/ICEAST.2018.8434395.

[29] A. Hussein, A. Wael, M. James, Collins G. Ntim, and W. Yan, "Prediction of Financial Strength Ratings Using Machine Learning and Conventional Techniques," *Invest. Manag. Financ. Innov.*, vol. 14, no. 4, pp. 194-211, 2017. doi: 10.21511/imfi.14(4).2017.16.

[30] I. C. Yeh and C. hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 2473–2480, 2009, doi: 10.1016/j.eswa.2007.12.020.

[31] A. G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, "A customized classification algorithm for credit card fraud detection," *Eng. Appl. Artif. Intell.*, vol. 72, pp. 21–29, Jun. 2018, doi: 10.1016/J.ENGAPPAI.2018.03.011.

[32] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," *IEEE Access*, vol. 6, pp. 14277–14284, Feb. 2018, doi: 10.1109/ACCESS.2018.2806420.

[33] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods," *Knowledge-Based Syst.*, vol. 128, pp. 139-152, 2017, doi: 10.1016/j.knosys.2017.05.001.

[34] H. H. Sun Yin, K. Langenheldt, M. Harlev, R. R. Mukkamala, and R. Vatrapu, "Regulating Cryptocurrencies: A Supervised Machine Learning Approach to De-Anonymizing the Bitcoin Blockchain," *J. Manag. Inf. Syst.*, vol. 36, no. 1, pp. 37–73, Jan. 2019, doi: 10.1080/07421222.2018.1550550.

[35] P. Hájek and V. Olej, "Concurrent Sentiment and Concept Extraction from Corporate Annual Reports for Financial Performance Forecasting," in *The 6th International Multi-Conference on Complexity, Informatics and Cybernetics*, 2015.

[36] E. Badal-Valero, J. A. Alvarez-Jareño, and J. M. Pavía, "Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case," *Forensic Sci. Int.*, vol. 282, pp. 24–34, 2018, doi: 10.1016/j.forsciint.2017.11.008.

[37] M. Jin, H. Wang, Q. Zhang, and C. Luo, "Financial Management and Decision Based on Decision Tree Algorithm," *Wirel. Pers. Commun. 2018 1024*, vol. 102, no. 4, pp. 2869–2884, Feb. 2018, doi: 10.1007/S11277-018-5312-6.

[38] Y. Li, C. Yan, W. Liu, and M. Li, "Research and application of random forest model in mining automobile insurance fraud," in *2016 12th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov. ICNC-FSKD 2016*, no. 61502280, pp. 1756–1761, 2016, doi: 10.1109/FSKD.2016.7603443.

[39] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud : A comparative study," *Decis. Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011, doi: 10.1016/j.dss.2010.08.008.

[40] R. A. Bauder and T. M. Khoshgoftaar, "Medicare fraud detection using random forest with class imbalanced big data," in *Proc. - 2018 IEEE 19th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2018*, pp. 80–87, Aug. 2018, doi: 10.1109/IRI.2018.00019.

[41] S. Ozdemir, *Principles of Data Science*. Packt Publishing Ltd, 2016.

[42] D. Lavanya and D. K. U. Rani, "Analysis of Feature Selection with Classification : Breast Cancer Datasets," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 5, pp. 756–763, 2011.

[43] "Audit Data Data Set." UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Audit+Data (accessed Jun. 02, 2020).

[44] N. A. B. Kamisan, S. M. Norrulashikin, and S. F. Hassan, "Missing Values Imputation For Wind Speed," *Appl. Math. Comput. Intell.*, vol. 10, no. 1, pp. 319–327, 2021.

[45] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowledge-Based Syst.*, vol. 41, pp. 16–25, Mar. 2013, doi: 10.1016/J.KNOSYS.2012.12.007.

[46] N. Hooda, S. Bawa, and P. S. Rana, "Fraudulent Firm Classification: A Case Study of an External Audit," *Appl. Artif. Intell.*, vol. 32, no. 1, pp. 48–64, Jan. 2018, doi: 10.1080/08839514.2018.1451032.

[47] P. Hájek and V. Olej, "Word categorization of corporate annual reports for bankruptcy prediction by machine learning methods," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9302, pp. 122–130, 2015, doi: 10.1007/978-3-319-24033-6_14.

[48] J. Brownlee, "Imbalanced Classification with Python Choose Better Metrics , Balance Skewed Classes , and Apply Cost-Sensitive Learning," *Machine Learning Mastery*.

[49] L. Van Der Maaten, E. Postma, and J. Van Den Herik, "Dimensionality Reduction: A Comparative Review," *J Mach Learn Res*, 2009.

[50] P. Harrington, *Machine Learning in Action*. Manning Publications Co., 2012.

[51] I. Sadgali, N. Sael, and F. Benabbou, "Performance of machine learning techniques in the detection of financial frauds," *Procedia Comput. Sci.*, vol. 148, pp. 45–54, Jan. 2019, doi: 10.1016/J.PROCS.2019.01.007.

[52] Y. Zhang and P. Trubey, "Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection," *Comput. Econ.*, vol. 54, no. 3, pp. 1043–1063, 2019, doi: 10.1007/s10614-018-9864-z.