Int. J. Nonlinear Anal. Appl. 15 (2024) 8, 259–272
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.30713.4471

Predicting fraud in financial statements using supervised methods: An analytical comparison

Zahra Nemati[a], Ali Mohammadi[a,*], Ali Bayat[a], Abbas Mirzaei[b]

[a] *Department of Accounting, Zanjan Branch, Islamic Azad University, Zanjan, Iran*

[b] *Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran*

(Communicated by Javad Vahidi)

Abstract

The current era is known as the "age of information," and the capital market is built on information as the economy's primary engine. The system of financial statements of corporations, which is the most significant source of information used in the capital market, produces an information system called accounting. Fraud and manipulation in these financial statements raise corporate risk, erode investor confidence, and cast doubt on the objectivity of accounting experts. Owing to the significance of fraud, this study aims to offer a way to foretell the likelihood of fraud in the financial statements of businesses admitted to the Tehran Stock Exchange between 2014 and 2021. 180 enterprises listed on the stock exchange make up the statistical sample (532 years of companies - suspected fraud years and 908 years - of non-fraudulent companies). According to the independent auditor's assessment, the existence of dormant assets and items, the doubting of the assumption of continuity of activity, the presence of tax discrepancies with other tax areas, and the dearth of adequate performance tax reserves led to the selection of the companies suspected of fraud. 96 financial ratios have been compiled by examining the theoretical foundations and research. In this research, the supervised methods of support vector machine, K-nearest neighbor, Bayesian network, neural network, decision tree, logistic regression, random forest and the hybrid method (bagging) have been used. The results of the research showed that the performance evaluation criteria of precision, accuracy, sensitivity, and F-Measure and efficiency (ROC) and the accuracy result of the confusion matrix in the combined method (bagging) were 72.45, 61.21, 64.74, 62.93, 73.50, and 72.45 percent, respectively, which indicates the better performance and greater ability of this method to predict the possibility of fraud in financial statements compared to other proposed methods.

Keywords: Fraud, Financial statements, Financial ratios, Supervised methods
2020 MSC: 91G15

1 Introduction

In recent decades, executives and law enforcement organizations around the world have focused heavily on the subject of misleading financial reporting. as stated in [23]. Since 2009, there have been startlingly frightening allegations of financial fraud being committed worldwide. According to the 2018 Global Economic and Fraud Survey, 49 percent of international organizations fell victim to financial crimes in 2017 and 2016 [23]. The average loss per fraud

[*] Corresponding author

Email addresses: znemati5879@gmail.com (Zahra Nemati), ali_mohammadi93@yahoo.com (Ali Mohammadi), ali.Bayat22@yahoo.com (Ali Bayat), a.mirzaei@iauardabil.ac.ir (Abbas Mirzaei)

Received: April 2023 *Accepted:* June 2023

is $593,000, making financial statement fraud one of the most costly types of fraud, although it accounts for only 9% of all incidents of fraud in reports. Fraud has dominated discussions of financial markets and domestic economic-social institutions for the past 20 years [20]. As a result, there are official organizations in the majority of industrialized nations that release statistics on the prevalence of fraud and list the names of fraudulent businesses, such as the [16]. They also passed laws like the Australian Companies Act of 2001, the Sarbanes-Oxley Act of the United States in 2002, and the United Kingdom Public Interest Disclosure Act of 1999 in an attempt to stop fraud. But despite the importance of fraud in financial statements and the fact that Iran is ranked 50th out of 180 countries in the world when it comes to the spread of financial corruption and 25th in Transparency International's ranking of financial corruption in 2021, there is currently neither a legal institution that can directly investigate and find fraud nor a database that can give information on the list of fraudulent companies to help identify them. On the one hand, with the advancement of technology and high-speed communication networks, nowadays the methods of committing fraud are very complex; committing it is easier, and detecting it has become more difficult, especially since fraud is usually hidden and fraudsters act smartly and quickly [25]. For this reason, its diagnosis is a complex but important task. So far, many models have been put forward to help find fraud. These models can be roughly put into three groups: statistical models, innovative models, and models based on artificial intelligence. So far, many models have been put forward to help find fraud. These models can be roughly put into three groups: statistical models, innovative models, and models based on artificial intelligence. Studies show that the use of statistical models, due to their reliance on limiting assumptions such as normal distribution and a high classification error rate, and meta-heuristic models, due to their being stuck in local optimal points and premature convergence, has decreased and led to the acceptance of artificial intelligence models because these techniques are frequently non-parametric and require little initialization. Because of this, artificial intelligence techniques are becoming more prevalent in fraud detection methods. As a result, we are searching for the best technique in this research to recognize fraudulent businesses using supervised methods that are artificial intelligence techniques in light of the aforementioned cases [28].

2 Theoretical Foundations and Research background

2.1 Definition of Fraud

Professional references and researchers have given different definitions of fraud, its risk factors, and its major causes.

1. Managers or other employees commit fraud as a result of a specific motivation or under a specific force.
2. A certain status quo such as the lack of control, inappropriateness of existing controls, or a manager's power to violate controls can provide an opportunity for fraud.
3. Those who commit fraudulent actions are able to justify their actions. Some of them have specific attitudes, characteristics, or moral principles allowing them to commit certain crimes intentionally [21].

Krancher [13] defined fraud any kinds of crime committed mainly by deception. They introduced four major elements of fraud as below in accordance with the customary law:

1. False important statement
2. Awareness of the falsity of a statement
3. Reliance (trust) of the deceived on a false statement
4. The loss of the deceived as a result of reliance on a false statement

2.1.1 Theoretical foundations

Financial fraud has a variety of definitions from a variety of perspectives.

Legal definition: Fraud is any illegal action involving deception, secrecy, and malversation requiring no physical violence or force. Fraud is committed in organizations to acquire money, assets, or service for commercial and personal benefit.

Fraud, in law, is any deceptive financial action by people or the simultaneous occurrence of the following elements:

Legal element (legislation for the financial crime)

Objective element (occurrence of financial crime)

Subjective element (criminal intent)

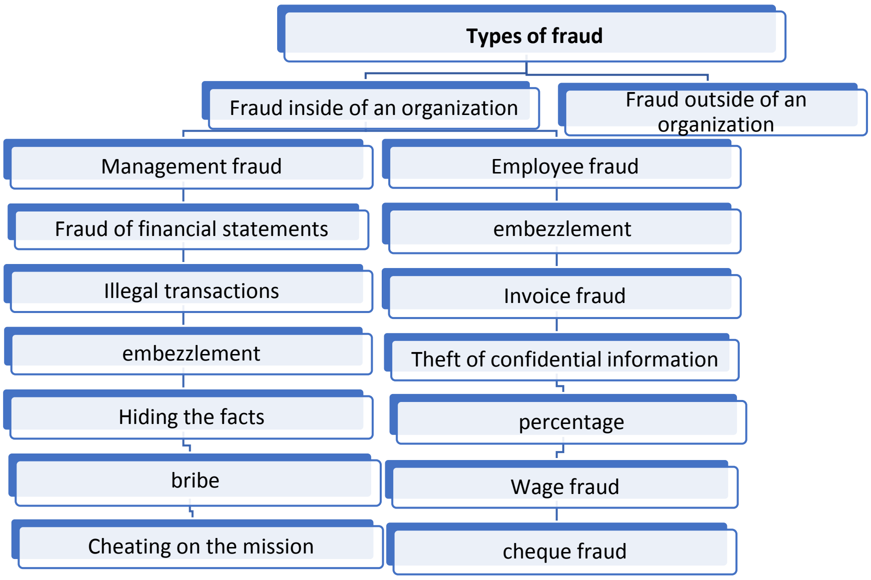

Figure 1: Types of fraud according to people who commit fraud [6].

2.2 Types of fraud

In a broad classification, internal and external fraud types can be separated into two categories:

Fraud inside of an organization: Both management and employees in the organization are involved in this scam.

Fraud outside of an organization: external individuals stealing from or abusing the organization's resources [12].

2.3 Suggested Algorithms:

Support-vector machine algorithm: One of the supervised learning classification techniques that can be utilized to address classification or regression issues is the support vector machine. Based on the idea of statistical learning and structural risk minimization, Vapnik (1995) developed this algorithm. The best super page is the one that has the greatest distance from two groups. This algorithm creates hyperplanes in space that perform the task of differentiating various data samples optimally, that is, it separates two groups in such a way that they have the greatest distance from each other's closest points. This method classifies the data by finding the best hyper maps that separate all the data of one group from the data of another group [18].

K-nearest neighbor algorithm: One of the most straightforward and crucial approaches for classifying data is this algorithm. The foundation of this algorithm is to identify a certain number of closest elements in the statistical community to the new element inserted in that community, based on which the closest existing data to the new element can be identified. Put it in the same class as elements that are similar to it based on various attributes. This approach, according to Kuncheva [14], is a kind of non-parametric classification used to extract the distribution function from dispersed data. It contains a document or training data for classification, and the algorithm determines similarities between the pre-classified training documents based on a criterion. These categories are then used to predict the category of the test document by scoring the documents in each chosen category. (Goa et al., 2002) In general, k nearest neighbor is a special case of instance-based learning that deals with symbolic data. This method is also an example of lazy learning, a technique that waits for the query to generalize beyond the training data [2].

Bayesian network algorithm: The English priest Thomas Bayes' discovery of the Bayes formula in 1763 is the beginning of the history of Bayesian networks. The likelihood of belonging to a particular group is estimated by this procedure, which is based on Bayes' probability theorem. Here is how Bayes' Theorem is put into words [15]:

$$P(Y|X) = \frac{P(Y) * P(X|Y)}{P(X)}$$

X is the observation (or set of attributes) and Y is the result (or group label) that can be obtained from the dataset.

$P(Y|X)$ represents the posterior probabilities of variable X over possible classes.

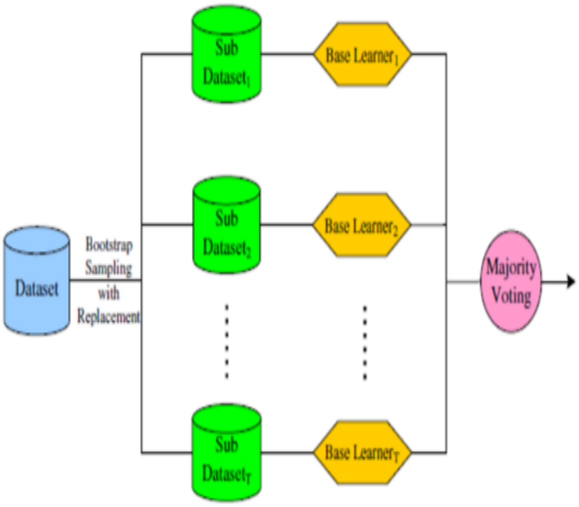

Figure 2: Algorithm of the combined method (Bagging) [30]

$P(Y)$ represents the prior probabilities of each class without information about variable X.

$P(X|Y)$ represents the conditional probabilities of the variable X with the probability of Y.

$P(X)$ is basically the probability of the evidence.

A new sample can be classified by obtaining $P(Y|X)$ for each distinct group of Y and determining which of these groups has a larger value. The estimated group for the new sample is the label of the particular group Y with the highest value of $P(Y|X)$ for a certain trait X. $P(X)$ does not have to be calculated for each new sample and is taken to be constant because it yields the same results for each specific group value [26].

The combined method's algorithm (Bagging): Berryman first developed the idea of collective learning in 1996, and the goal of this algorithm is to minimize error by combining a number of related machine learning models. Each classifier method in the bagging approach creates a model using the training data so that it can recognize differences between various classes. By using the models created by the other classification techniques instead of creating its own, this algorithm chooses the class for the current sample based on a vote. The data set can be accessed by each classifier. Each classifier in this method receives a portion of the original data set. This means that each classifier must base its model on the same portion of the data that has been handed to it and that it sees in the data set [26].

Neural network algorithm: This algorithm's design was inspired by the way the human brain is organized. Neural networks may be trained to mimic the behavior of the human brain to recognize patterns and classify data in a manner similar to how the human brain recognizes various patterns of data and categorizes different sorts of information. The number of nerve cells and other similar processing units acting as interconnected processing units make up a neural network algorithm [8]. This technique is insensitive to outlier data and is effective for categorizing and predicting variables that are challenging to measure, as well as for addressing non-linear classification problems. It links inputs through non-linear information processing to outputs that are matched to goals in a network of artificial neurons that create layers of hidden units [4]. The activity of each hidden unit and the output Y are calculated from the combination of the input X and the set of neuron weights W:

$$Y = f(X, W)$$

Decision tree algorithm: Based on the input data, this algorithm uses a tree structure to predict the outcome or attribute value. Because a set of key rules may be achieved by matching the requirements from the root node to the leaf nodes, this method is a common tool in machine learning that aids in finding the best approach to achieving a good result. As the name implies, classification trees are used to divide a data set into classes that belong to the response variable. The response variable often has a two-category yes or no format (1 or 0). Different decision tree methods are employed if the answer variable contains more than two categories [2].

Logistic regression algorithm: This procedure, which is a statistical technique for categorizing binary data, uses a linear model called a logistic or logit model to conduct regression on a set of variables. This approach uses a variety of modeling and analysis approaches to examine particular and distinctive factors, concentrating on the correlation between the dependent variable and one or more independent variables. Similar to linear regression, the transformed variable in the logit technique is approximated using a linear function. The weights must be chosen to produce the correct training data, as seen in linear regression. The error square approach is used in linear regression to assess the goodness of fit. The logistic regression technique use the following equation:

$$\sum_{n=1}^{\infty}(1-x^i)\log\big(1-p_r[1\mid a_1+a_2+\cdots+a_40\big)+x^i\log\big(pr[a_1^i+a_2^i+\cdots+a_k^]\big)$$

That x^i is equal to zero or one and they prefer and use it more [19].

Random forest algorithm: The output of the random forest method is based on the individual votes of each of the decision trees, and the final classification is completed. The random forest algorithm uses a huge number of decision trees. This approach generates a random data set and trains each decision tree using sampling with replacement. Entropy and the Gain Information criterion are used by RF to determine which feature is most crucial in each node of a tree. Gain Information is a technique for identifying the most informative characteristic, while Entropy assesses the impurity or uncertainty of attributes. As a result, when choosing characteristics, the objective is to reduce entropy and increase information gain. Additionally, RF chooses at random the subspace at each node of the tree, where a subset of m features is taken into account for decision-making at that node [1].

2.4 Research Background

This section will discuss a few of the domestic studies that are linked because the main focus of the essay is on the identification of fraud in the Iran Stock Exchange Organization.

In order to extract financial indicators from financial statements and textual features from management reports, Shugu and Shen Yang [29] introduced a financial fraud detection system that uses a deep learning model based on a mix of numerical features and textual data. To assess the performance of deep learning models using numerical data, textual data, and a mix of the three, they used the annual reports of 5130 Chinese corporations in the MD & A section. The outcomes demonstrated that the suggested strategy, which uses the LSTM classifier with 94.98% accuracy and the GRU classifier with 94.48% accuracy are better than traditional classification methods.

To find fraud in financial statements, Gupta and Mehta [7] used two machine-learning approaches and statistical methods. The results of the investigation show that the accuracy of the logistic regression method is 71.5%, perbit regression is 89.5%, the neural network is 71.7%, the decision tree is 73.6%, and the support vector machine is 90.4%. And the fuzzy method has an accuracy of 86.8 percent. which demonstrates the effectiveness of machine learning techniques over statistical approaches in identifying corporate fraud, especially when the sample size is small. Hedayatullah [9]. In order to detect fraudulent financial statements in Indonesian companies using machine learning based on meta-heuristic optimization, they initially selected 18 financial ratios for which information was available, and then using principal component analysis, 10 financial ratios were extracted, which are used in classification.

Then, they developed financial statement fraud prediction models using a variety of machine learning techniques based on meta-heuristic optimization. In this study, financial variables have been reduced using genetic algorithms, support vector machines, and improved back propagation neural networks for classification. The accuracy of financial ratios retrieved using genetic algorithms and vector machine classification is higher than that of other techniques (96.15%).

Sadgali [25], in a research titled "The performance of machine learning models in fraud detection,“ investigated data mining methods to detect fraud. And the results show that the probabilistic neural network with a number of 98.09% was the most accurate among the other methods.

Yao [30] suggested a model to identify financial statement fraud using data mining techniques. 17 financial ratios extracted with the two methods of principal component analysis and Xgboost are given to 6 and 5 financial reduction ratios, respectively. After that, they classified businesses into fraudsters and non-fraudsters using five classification techniques based on the financial ratios they had extracted: logistic regression, support vector machine, random forest, decision tree, and artificial neural network. The results of this research showed that the support vector machine method, with 71.67% accuracy, had the best accuracy and random forest, with 68.17% accuracy, had the lowest accuracy among the proposed methods. Johari and Smith [17] used logistic regression techniques, support vector machines, multi-criteria decision models, and artificial neural networks to study the prediction of fraud in financial reporting. The 10 best financial ratios are utilized in this study to forecast fraud. The findings demonstrated that the artificial neural network approach has a higher ability to anticipate fraud than other methods, with a prediction level of 94.87%.

Using the snowball approach, the study of literature, and expert interviews, [10] retrieved two non-financial ratios and 19 financial ratios. They also predicted and detected fraud using support vector machines and neural network approaches. The results showed that the prediction power of the support vector machine is 86% and that it performs better than the neural network.

Alghiani [31] predicted the risk of fraud in tax financial reporting using a hybrid approach of traditional data mining, ANFIS, and a meta-heuristic algorithm. By utilizing several optimization algorithms in the data mining approach, the results demonstrate an increase in the predictive power of the financial-tax reporting identification model. The optimization with the particle swarm algorithm has produced the most optimal model.

To identify fraud, Rezaei [24] used 41 financial and non-financial data and 5 approaches, including Bayesian networks, decision trees, neural networks, support vector machines, and a hybrid approach. The study's findings suggested that the combined method, which has a 96.2% prediction rate, has higher evaluation power and accuracy than other methods.

Tashdidi [27] chose 23 financial ratios whose information was available in Iran in order to offer a fresh method for forecasting the detection of fraud in financial statements. Then, 16 ratios were determined to be the best and most efficient ratios using the mutual entropy approach. To categorize companies into fraudsters and non-fraudsters, three techniques are used: logistic regression, genetic algorithms, and honey bees. The findings demonstrate that, with a prediction accuracy of 82.5%, the honey bee algorithm is superior to other techniques in identifying fraud.

In order to choose the characteristics that have the biggest influence on detecting fraud in financial statements, Ebrahimi and Khajawi [5] employed a correlation-based feature selection method. They employed 40 financial and nonfinancial variables to do this, and the results show the value of cash financial ratios, interest coverage, accounts receivable to total assets, inventory to net sales, the natural logarithm of sales, net profit to sales, and current assets to total assets. In order to anticipate fraud, they also employed artificial neural networks, Bayesian networks, and random forest data mining approaches. The random forest algorithm outperformed the other methods with 96.77% accuracy.

A model to forecast the likelihood of financial fraud was proposed by Zare Behnamiri and Malekian in [32]. Seven ratios—working capital to assets, accounts receivable to sales, cash to current liabilities, balance to current assets, debt to equity, gross profit to assets, and the absolute value of changes in the current ratio-were examined using a MATLAB software method and in two stages using stepwise regression and elastic net tests. As a result of the study, the estimated model has a predictive ability of 64.04% when tested using the logit method.

Kazemi [11] employed data mining techniques to find instances of fraud risk in financial accounts, including logistic regression, the artificial neural network algorithm k-scale, and meta-heuristic techniques such as algorithms based on distance and entropy, and evolutionary algorithms. He investigated each of the aforementioned methods in 82 Iranian businesses. Comparing other methodologies, the distance-based Ants algorithm performed better.

3 Research hypothesis

In order to predict the possibility of fraud in financial statements, the combined method (Bagging) has a better performance than other monitored methods.

4 Research method

This research in terms of results and practical consequences, in terms of purpose, because it seeks to describe the conditions or phenomena under investigation "descriptive-correlation“. In terms of the implementation process, it is quantitative and in terms of time position, it is retrospective (post-event). In order to collect the required theoretical and background information, it has been done by studying the researches done inside and outside the country, including books and articles, using the library method and scanning. Excel software was utilized for their calculations in order to get the necessary data of the variables from the financial statements and reports of independent auditors and legal inspectors published by the Tehran Stock Exchange Organization and Rahvard Navin software package. Meta-innovative and data mining techniques are utilized to evaluate the information and test the research's hypotheses, and for this, MATLAB and Datalab tools are used.

5 Statistical population and research sample

The companies admitted to the Tehran Stock Exchange between 2014 and 2021 make up the statistical population for the current study. Companies that meet all of the criteria listed below are examples of those accepted by the Tehran Stock Exchange:

1. Companies that have been admitted to the stock exchange by the end of 2013 and the name of the company has not been removed from the Tehran Stock Exchange during the period under review.
2. Their financial year should end at the end of March and it has not changed in the studied period of the financial year.
3. Do not be a part of financial intermediary businesses including holding, investing, banks, and insurance firms. due to the fact that these businesses are different in type and classification from production and service organizations in terms of financial statement items, operational activities, and financial provision.
4. The required information including financial statements and independent auditor's report for the research period should be available. According to the considered conditions, 180 companies were selected.

6 Research variables

Dependent variable: In order to define and specify fraud in financial statements as a dependent variable, examples of fraud were identified and extracted by consulting and studying Auditing Standard 240, titled "Auditor's Responsibility and Theoretical Foundations of Internal and External Investigations Related to Fraud." [3]

The following conditions are examples of those that may indicate fraud in Appendix Three- Auditing Standard 240:

- Transactions that are not recorded completely or on time, or the amount, registration date, classification, or accounting procedure, are inappropriate.
- Lack of documentation: not having enough documentation to determine the impact of possible adjustments.
- Existence of significant items that are unexplained in the form of disparities, unusual changes to the balance sheet, changes in trends, or significant ratios or linkages between financial statements, such as: rapid growth of receivables compared to the growth of income
- The financial results have been significantly impacted by the most recent adjustments.

Additionally, by examining both domestic and international research on fraud, the following fraud cases have been identified as being the most significant:

- Overstating and understating incomes and assets
- Overstating and understating in expenses and debts
- Updated financial statements and significant annual adjustments.
- the existence of tax disparities between tax areas and a lack of performance tax reserves
- The existence of assets and stagnant inventory items
- Consider a scenario in which the auditor's opinion is qualified and the consistency of the company's activity has been questioned for a series of consecutive periods, yet the company continues to prepare its financial statements based on the consistency of activity. For instance, a business that has had no sales since its production was halted two years ago
- Misapplication of accounting standards related to identification, measurement, classification, presentation, and disclosure

Because there is no organization or institution in Iran to identify companies that commit fraud and information about these companies is not available to the public, and on the other hand, considering that in some research, the relationship between fraud and The auditor's opinion is confirmed, The condition clauses and other clauses of the audit reports of the companies that have a modified opinion (rejected opinion, no opinion, and conditional opinion) are consequently thoroughly examined in light of the examples of fraud previously mentioned and in a sample of 1440 companies (180 companies for 8 years), 532 were suspected of financial statement fraud, while 908 were found to be non-fraudulent. Companies suspected of fraud are represented by the number one, while non-fraudulent companies are represented by the number zero.

Independent variable: financial ratios were used as independent variables or predictors of financial statement fraud in this study. The financial ratios were extracted into the four categories of liquidity, leverage, efficiency, and profitability by examining the theoretical foundations and research, and some of the ratios that were similar and opposite to each other were removed in the initial review. Finally, the following 96 financial ratios remain.

7 Research findings

The supervised methods used in this study include support vector machines, K-nearest neighbors, simple Bayesian models, neural networks, decision trees, logistic regression, random forests, and group classification using 96 financial ratios extracted from financial statements. The likelihood of fraud has been predicted using sample companies.

Learning techniques are first taught in order to examine and measure the ratios extracted from the proposed algorithms. In order to determine the training percentage of the models, 70% of the data (1008 data, of which 376 are suspected fraud companies and 632 are non-fraudulent companies) were provided as training data in the MATLAB software. Lastly, 30% of the remaining data (including 432 data, 156 data of companies suspected of fraud, and 276 data of non-fraudulent companies) were used as test data in the software to evaluate the algorithms and the expected amount of fraud in financial statements that should be investigated and evaluated.

Standards for evaluating the effectiveness and capability of the suggested techniques: The following definitions of the accuracy, precision, recall, and F measure can be used to evaluate how well each method for predicting the possibility of financial statement fraud performs:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{Recall} &= \frac{TP}{TP+FN} \\
F\text{-measure} &= \frac{2 * \text{Precision} * \text{Recall}}{Precision + Recall}
\end{aligned}$$

According to the financial ratios extrapolated from the sample companies' financial statements, the results of these performance evaluation criteria for each of the proposed methods are as follows.

In Figure 3, the outcome of analyzing the accuracy criteria for each of the suggested methods is displayed.

As shown in the aforementioned figure, among the classification methods, the combined method (Bagging) with a number of 72.45% shows a better accuracy than other proposed methods. The support vector machine method with a number of 69.44% got the second place and the k-nearest neighbor method with a number of 66.20% got the third place.

The comparison of the Precision criteria for each of the suggested methods is shown in Figure 4.

According to this figure, it can be seen that among the classification methods, the combined method (Bagging) with a number of 61.21% shows a better F-measure than other proposed methods. The support vector machine method with a number of 56.90% got the second place and the k-nearest neighbor method with a number of 52.72% got the third place.

Figure 5 displays the outcome of the performance evaluation for each of the suggested methods.

Among the classification methods, the combined method (Bagging) with a value of 64.74% shows better sensitivity than other classification methods. The support vector machine method with a number of 63.46% has obtained the second rank and the k-nearest neighbor method has obtained the third rank with a number of 62.18%

Figure 6 shows the comparison of the F-measure in each of the proposed methods.

According to this figure, it can be seen that among the classification methods, the combined method (bagging) with a number of 62.93% shows a better F-measure than other proposed methods. The support vector machine method

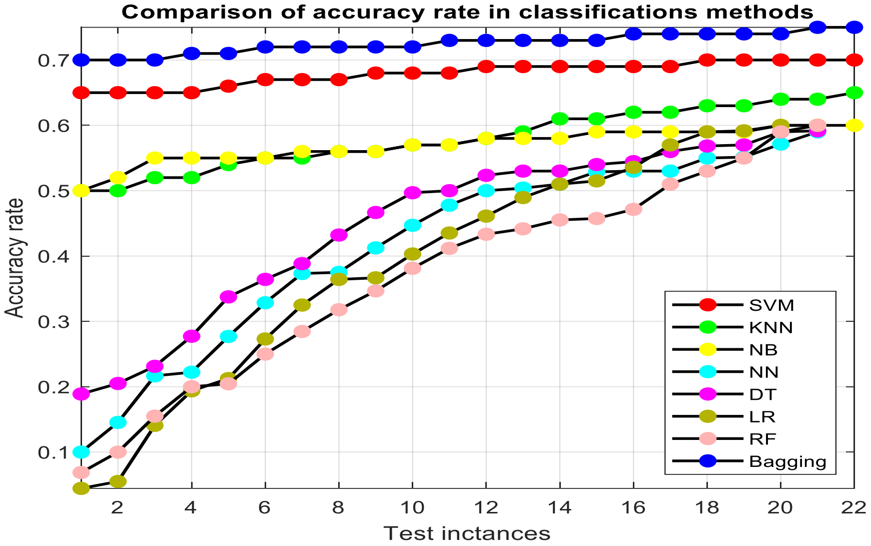

Figure 3: Comparison of accuracy criteria for the proposed methods

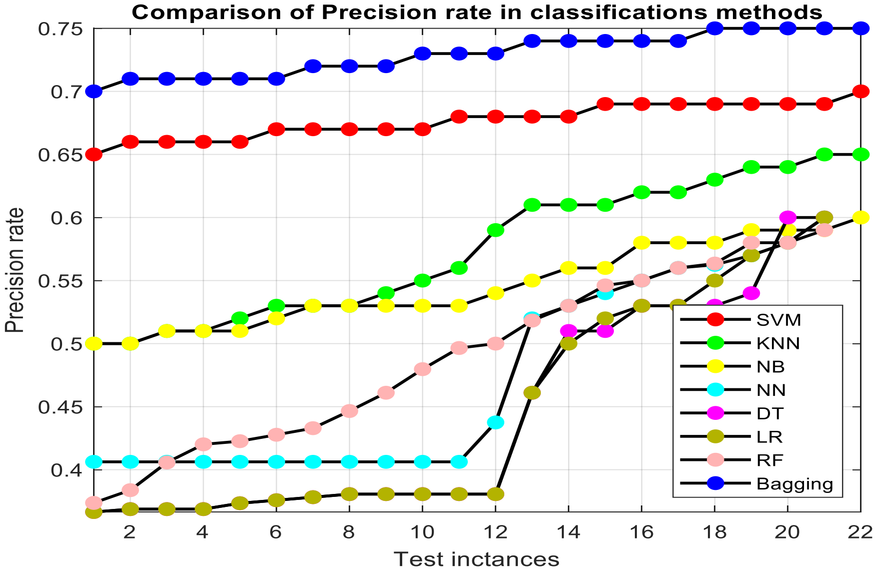

Figure 4: Comparison of Precision criteria for the proposed methods

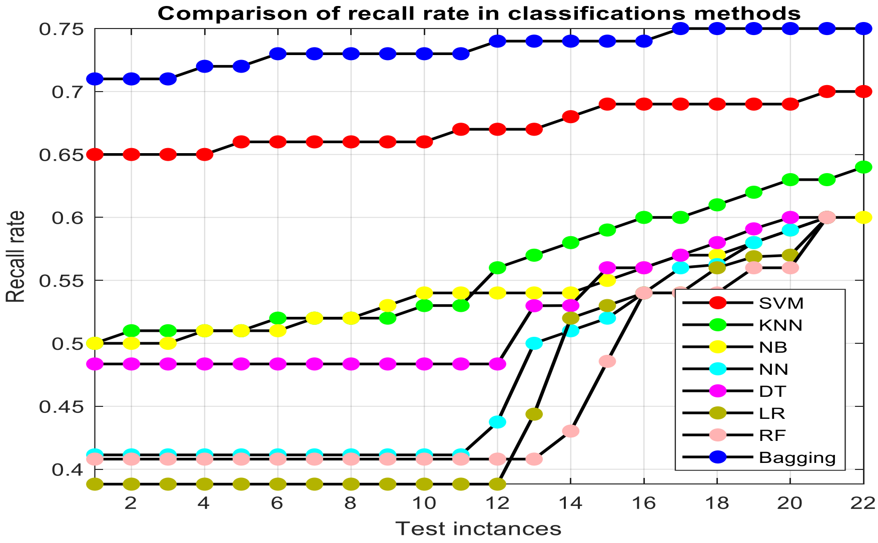

Figure 5: Comparison of the Recall criteria for the proposed methods

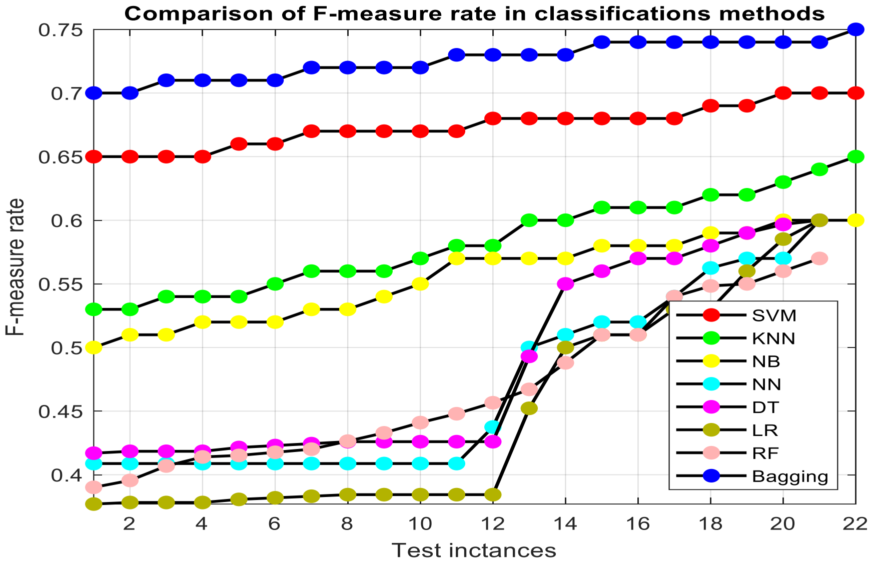

Figure 6: Comparison of the F-measure for the proposed methods

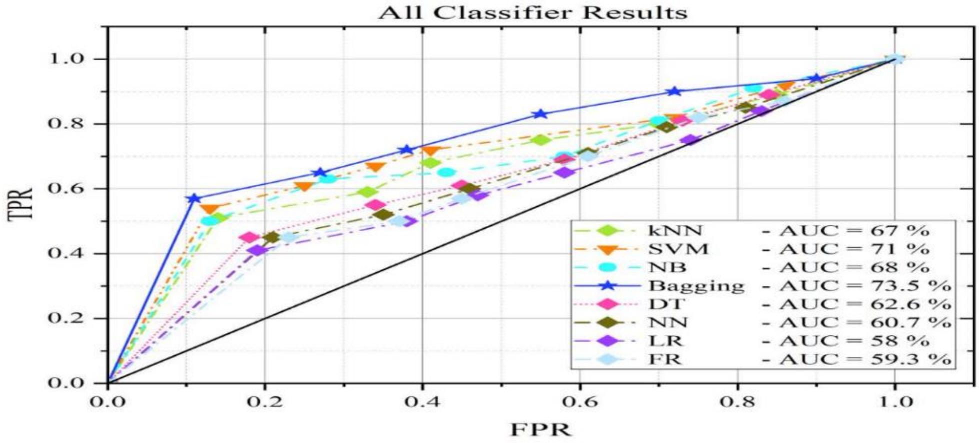

Figure 7: Measuring and evaluating the effectiveness of the proposed classification methods with 96 financial ratios.

with a number of 60% got the second place and the k-nearest neighbor method with a number of 57.06% got the third place.

2-7) Examining the confusion matrix of the proposed methods for predicting the possibility of fraud in financial statements:

The number of rows and columns of the confusion matrix depends on the number of classes, which in this research has two classes of non-fraudulent and suspected fraud companies, therefore, the confusion matrix in this research is:

True Positive (TP): Financial statements that are suspected of being fraudulent and have been correctly identified.

False Positive (FP): financial statements that are suspected of being fraudulent and have been wrongly identified as non-fraudulent.

True Negative (TN): financial statements that are not fraudulent and are correctly identified.

False negative (FN): financial statements that are not fraudulent and have been mistakenly suspected of fraud.

Table 1: The results of the confusion matrix of the proposed methods

Description	False negative (FN)	True negative (TN)	False positive (FP)	True positive (TP)	Prediction accuracy
Combined method (Bagging)	55	212	64	101	72.45%
Support vector machine	57	201	75	99	69.44%
k nearest neighbor	59	189	87	97	66.20%
Bayesian network	64	191	85	92	65.51%
neural network	85	164	112	71	54.40%
decision tree	94	172	104	62	54.17%
logistic regression	100	151	125	56	47.92%
Random forest	107	132	144	49	41.90%

7.3) The receiver operating characteristic (ROC) of the proposed methods for predicting the possibility of fraud in financial statements:

The ROC curve is a two-dimensional drawing of the performance of the proposed methods, in which the true positive is drawn on the X graph and the false positive is drawn on the Y graph. A common criterion in this method is to calculate the area under the ROC chart. The ROC chart below illustrates how well the proposed classification algorithms predict the likelihood of fraud in financial accounts based on the accuracy, sensitivity, accuracy, true positive rate, and false positive rate data.

8 Conclusion

Given that many of their users, including shareholders, creditors, investors, etc., rely on financial statements to make decisions, Therefore, the presence of fraud in these cases not only results in expense, risk, and errors in their decisions but also has serious non-financial effects, particularly the erosion of the credibility of the accounting profession. On the other hand, as information technology has developed, fraudsters' methods have become faster and more complex, and it is unclear exactly what independent factors affect the accuracy of financial statement fraud predictions. Compared to discrete and non-linear data, continuous and linear data are more predictive using statistical methods. The likelihood of accurately predicting fraud is decreased with these methods [22]. The development of fraud detection techniques in financial statements is crucial for this reason. In order to determine which supervised classification method performs best, the current study compares a number of them, including support vector machines, K-nearest neighbors, Bayesian networks, neural networks, decision trees, logistic regression, random forests, and group classification. Additionally, using financial ratios, it is better able than other techniques to foresee the possibility of fraud in financial statements. According to the obtained results, the Bagging classification method performs better than other supervised classification methods for all performance evaluation criteria.

Table 1 shows the average values of the evaluation criteria for the proposed methods, therefore, according to the results of the research hypothesis that has been stated to predict the possibility of fraud in financial statements, the combined methods (bagging) are better than other monitored methods. It has better performance, it is confirmed.

Table 2: Brief results of performance evaluation criteria for proposed algorithms and confusion matrix

Description	Accuracy	Precision	Recall	F-measure	ROC	$TP+TN$
Combined method (bagging)	72/45%	61.21%	64.74%	62.93%	73.50%	313
Support vector machine	69.44%	56.90%	63.46%	60%	71%	300
k nearest neighbor	66.20%	52.72%	62.18%	57.06%	67%	286
Bayesian network	65.51%	51.98%	58.97%	55.26%	68%	283
neural network	54.39%	38.79%	45.51%	41.88%	60.70%	235
decision tree	54.16%	37.34%	39.74%	38.50%	62.60%	234
logistic regression	47.91%	30.93%	35.89%	33.22%	58%	207
Random forest	41.89%	25.38%	31.41%	31.60%	59.30%	181

Suggestions

According to the research results, the following suggestions can be made:

- Since the AI algorithms and intelligent techniques are very accurate and fast in prediction with respect to the large amounts of data, researchers are advised to use these methods in their studies in order to faster detect cases of fraud and impose less loss on stakeholders.
- The esteemed legislating organizations and institutions can reduce the number of fraud cases in financial statements by modifying the trade laws, embedding law-binding control tools, considering preventive punishment methods, and increasing the fines.
- It is difficult, specialised, and time-consuming to identify the companies committing fraud in their financial statements, and many users of financial statements lack the necessary and sufficient expertise in this regard. Hence, an organisation or institution should consider addressing fraudulent financial reporting more seriously than ever before. Moreover, a specialised council should be formed to identify fraudulent companies and to publicise their information.

References

[1] M. Belgiu and L. Drăguţ, *Random forest in remote sensing: A review of applications and future directions.* J. Photogram. Rem. Sen. **114** (2016), 24-31.

[2] C.B. Rjeily and G. Badr, A.H. El Hassani, and E. Andres, *Medical data mining for heart diseases and the future of sequential mining in medical field*, Machine Learning Paradigms, Springer, 2019, pp. 71–99.

[3] Committee of Audit Standards. *Standards and Principles of Accounting and Auditing: Audit Standards*, Audit Organisation Publications: Tehran, Iran, 2015.

[4] S. Dua and X. Du, *Data mining and machine learning in cybersecurity*, CRC Press, 2016.

[5] M. Ebrahimi and Sh. Khajavi, *Modeling variables affecting fraud detection in financial statements through data mining techniques*, Financ. Account. **33** (2017), 41–62.

[6] P. Goldmann and H. Kaufman, *Anti-Fraud Risk and Control Workbook*, John Wiley & Sons, 2009.

[7] S. Gupta and S.K. Mehta, *Data mining-based financial statement fraud detection: Systematic literature review and meta-analysis to estimate data sample mapping of fraudulent companies against non-fraudulent companies*, Glob. Bus. Rev. (2021), https://doi.org/10.1177/0972150920984857

[8] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edition, University. Ill. Urb. Mic. Kam. J. P. S. F. Uni., 2012.

[9] S. Hidayattullah, I. Surjandari, and E. Laoh, *Financial statement fraud detection in Indonesia listed companies using machine learning based on meta-heuristic optimization*, Int. Workshop Big Data In. Sec. (IWBIS). IEEE, 2020, pp. 79–84.

[10] H. Kamrani and B. Abedini, *Formulation of financial statement fraud detection model using artificial neural network and support vector machine approaches in companies listed in Tehran Stock Exchange*, J. Manag. Account. Audit. Knowl. **11** (2022), no. 41, 285–314.

[11] T. Kazemi, *Identifying cases of fraud risk in financial statements of Iran and evaluating fraud detection methods*, Doctoral thesis, 2016.

[12] A. Khorasani, *Investigating the effects of applying auditing standards in information disclosure in fraudulent financial reporting*, Doctoral thesis, 2017.

[13] M.J. Kranacher and R. Riley, *Forensic Accounting and Fraud Examination*, John Wiley & Sons, 2011.

[14] L.I. Kuncheva, *Combining pattern classifiers: Methods and algorithms*, John Wiley & Sons, 2014.

[15] K.M. Leung, *Naive Bayesian classifier*, Poly. Uni. Dep. Com. Science/Finance and Risk Engin. **2007** (2007), 123–156.

[16] Fraud, Occupational, *A Report to the nations, ACFE: https://acfepublic. s3. us-west-2. Amazonaws. com/2022+ Report+ to+ the+ Nations. pdf*, 8 (2023).

[17] N. Omar, Z.A. Johari, and M. Smith, *Predicting fraudulent financial reporting using artificial neural network*, J. Finan. Crime. **24** (2017), no. 2, 362–387.

[18] A. Pradhan, *Support vector machine-a survey*, Int. J. Emer. Tech. Adv. Engin. **2** (2012), no. 8, 82–85.

[19] Y. Park and D. Reeves, *Deriving common malware behavior through graph clustering*, In Proc. 6th ACM Symp. Info. Com. Commu. Sec., 2011, March, pp. 497–502.

[20] A.M.R. Nafchi and M. Dastgir, *Proposing a Model for Identification of Risk Factors Affecting Fraud in Banks.* Financ. Account. **41** (2018), 23-–45.

[21] M. Ramos, *Auditors responsibility for fraud detection*, J. Account. **195** (2003), no. 1, 28–36.

[22] P. Ranganathan, C.S. Pramesh, and R. Aggarwal, *Common pitfalls in statistical analysis: Logistic regression*, Perspect. Clinic. Res. **8** (2017), no. 3, 148.

[23] S. Rastatter, T. Moe, A. Gangopadhyay, and A. Weaver, *Abnormal Traffic Pattern Detection in Real-Time Financial Transactions*, Technical Report, EasyChair, 2019.

[24] M. Rezaei, M.N. Ardakani, and A.N. Sadrabadi, *Fraud detection in financial statements through audit reports of financial statements*, Manag. Account. **13** (2020), no. 45, 141—153.

[25] I. Sadgali, N. Sael, and F. Benabbou, *Performance of machine learning techniques in the detection of financial frauds*, Procedia Comput. Sci. **148** (2019), 45–54.

[26] A. Shinde, A. Sahu, D. Apley, and G. Runger, *Preimages for variation patterns from kernel PCA and bagging*,

Iie Trans. **46** (2014), no. 5, 429–456.

[27] E. Tashdidi, S. Sepasi, H. Etemadi, and A. Azar, *A novel approach to prediction and detection of fraud in financial statements through bee colony optimisation algorithm.* Account. Knowledge **12** (2019), 139-–167.

[28] W. Xiuguo and D. Shengyong, *An analysis on financial statement fraud detection for Chinese listed companies using deep learning*, IEEE Access **10** (2022), 22516–22532.

[29] J. Yao, Y. Pan, S. Yang, Y. Chen, and Y. Li, *Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: A multi-analytic approach*, Sustainability **11** (2019), no. 6, 1579.

[30] J. Yao, J. Zhang, and L. Wang, *A financial statement fraud detection model based on hybrid data mining methods*, Int. Con. Art. Int. Big Data (ICAIBD) IEEE., 2018, May, pp. 57–61.

[31] M.Y. Alghiani, J.S. Bahri, S.J. Kangarlouei, and A.Z. Rezaei, *Explaining financial tax cross reporting of companies: Hybrid method of classic data mining, ANFIS, and metaheuristic algorithms*, Empir. Stud. Financ. Account. **18** (2021), no. 71, 89-–111.

[32] M.J.Z. Bahmanmiri and E.M. Kalebastani, *Ranking the factors affecting financial fraud probability, according to audited financial statements*, J. Empir. Res. Account. **6** (2016), no. 3, 1–18.