

# Financial Ratios and Efficient Classification Algorithms for Fraud Risk Detection in Financial Statements

Z. Nemati<sup>1</sup>, A. Mohammadi\*<sup>1</sup>, A. Bayat<sup>1</sup>, A. Mirzaei<sup>2</sup>

<sup>1</sup>Department of Accounting, Zanzan Branch, Islamic Azad University, Zanzan, Iran,

<sup>2</sup>Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran.

Submission Date: 2023/04/11, Revised Date: 2023/05/10, Date of Acceptance: 2023/10/28

## Abstract

This study aims to identify the best financial ratios and the most efficient method for fraud risk detection in the financial statements of the listed companies by considering the financial importance of decision-making as well as the growing fraud statistics and detrimental effects. The statistical sample included 180 companies listed in the Tehran Stock Exchange from 2014 to 2021 (532 fiscal years suspected of fraud and 908 non-fraudulent fiscal years). Theoretical foundations were first taken into account to extract 96 financial ratios. The k-NN algorithm, Bayesian network, support vector machine, and bagging method were then employed for fraud risk detection in financial statements. According to the findings, the adopted methods failed to meet the evaluation standards in general. With an accuracy of 70.60% and a proportionality function value of 0.2940, the gray wolf optimization (GWO) algorithm was then utilized to reduce the ratios in order to improve performance. After 31 iterations, nine appropriate financial ratios were determined. The extracted financial ratios were then used to reevaluate the effectiveness of the proposed fraud detection strategies. After the financial ratios were reduced, all of the proposed approaches yielded better results. The accuracy and efficiency of the bagging method, support vector machine, Bayesian network, and k-NN algorithm were reported 79.25% and 81.70%, 75.83% and 80.30%, 72.01% and 74.60%, and 74.55% and 75.60%, respectively. In conclusion, the bagging method outperformed the other approaches in terms of accuracy and efficiency.

**Keywords:** Metaheuristic Algorithm, Data Mining, Financial Ratios, Classification Algorithms, Fraud Risk Detection.

---

\* Corresponding author: Email: ali\_mohammadi93@yahoo.com

## 1. Introduction

We are now living in the Information Age when the capital market is considered the driving force of economy based on information. In this age, accounting is an information system that produces financial statements of companies, which are used as the most important sources of information in a capital market. If this information is accurate, decent, clear, and reliable, it can greatly help users make investment decisions. In recent decades, fraudulent financial reporting has been among the most intriguing topics raised by law enforcers and legal institutions worldwide [1].

According to the Association of Certified Fraud Examiners (ACFE), although financial statement fraud accounts for only 9% of criminal cases, its average damage rate is \$593,000 per fraud, a figure which indicates the costliest case of financial crimes [2]. Prevention, detection, and investigation of financial statement fraud of companies have now become the new accounting concerns more than ever before. Nearly all organizations have somehow encountered different cases of fraud ranging from a negligible theft committed by an employee to fraudulent financial reporting. Major cases of fraud in financial statements can have substantially adverse effects on the market value of a business and its credibility and ability to achieve strategic goals, resulting in bankruptcy and loss of tens of thousands of job opportunities. In society, it can also damage the financial market efficiency, destroy the public trust in accounting and auditing, and harm economic developments [3].

All over the world, legislators have passed different laws to support the prevention of fraud. Instances are the UK Public Interest Disclosure Act 1999, the Australia Corporations Act 2001, and the US Sarbanes Oxley Act 2002. In most of the developed countries, there are also official organizations for reporting statistics regarding the occurrence of fraud and introducing fraudulent companies. The Association of Certified Fraud Examiners is an exemplar of such organizations in the US. Conducting an analysis of fraud on a global scale twice a year, this association detects cases of fraud and financial scandals and publishes a comprehensive report on various types of fraud, frequency of fraud, and financial impacts. Despite the importance of financial statement fraud risk detection in Iran, there are no legal institutions that can directly analyze and detect cases of financial fraud. Furthermore, there are no databases for disseminating the list of fraudulent companies. In fact, the cases of fraud investigated in the Tehran Stock Exchange are announced privately but not publicly if judicial courts reach and issue verdicts [4].

Pointing out an auditor's responsibility for detecting fraud and fault while auditing financial statements, Iran's Audit Standard 240 urges auditors to consider the concept of fraud in financial statements. However, according to *Section 4* of this standard, even if an audit is planned and implemented properly in accordance with relevant standards, fraud will probably be concealed. With advances in technology and high-speed communication networks, methods of fraud have become so sophisticated that it is now easier to commit fraud but more difficult to detect cases of fraud. In fact, fraudsters now act intelligently and quickly [5]. Hence, fraud detection is now a very difficult and complicated but important task. Thus, researchers have gradually started using artificial intelligence techniques rather than conventional methods and statistical analyses due to their reliance on restrictive hypotheses such as normal distribution and high classification error rates [6]. Given the importance of fraud risk detection in financial statements, this study aims to select appropriate financial ratios and adopt an efficient classification method for this purpose.

## 2. Theoretical Foundations and Research Background

### 2.1. Definition of Fraud

According to a pervasive definition by the ACFE (2012), fraud denotes the use of all various manmade tools by an individual to gain an advantage over another individual through false advice or concealment of the truth. In fact, fraud includes all abrupt events, tricks, deceptions, secrecy, and other unfair methods of cunning.

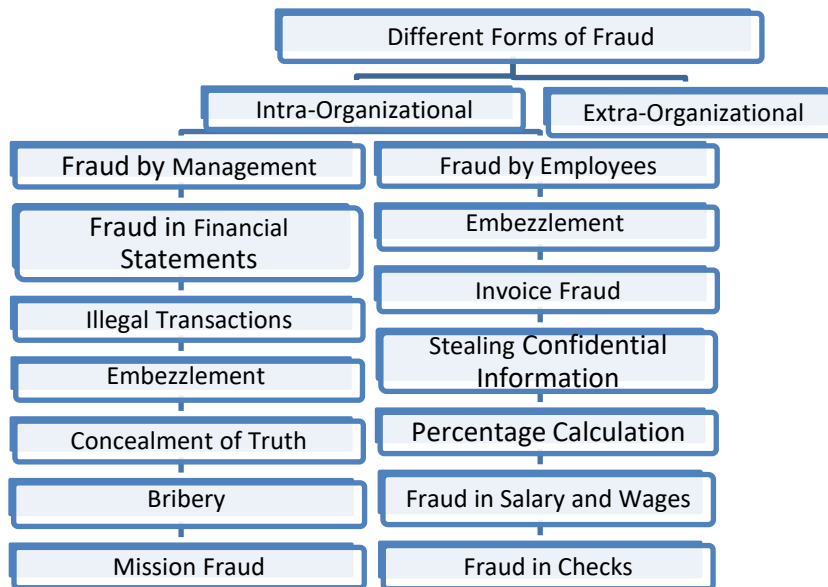
**Standard Accounting Definition of Fraud:** According to *Section 24* of Audit Standards, the distortion of financial statements can ensue from fraud or mistakes. Based on this standard, “fraud” denotes any deliberate or deceptive actions taken by one or several individuals such as managers, employees, or third parties to gain an illegal advantage. Although fraud is considered a broad legal concept, auditors are concerned about fraudulent actions leading to substantial distortion of financial statements [7].

### 2.2. Fraud Classification

In a general classification, different forms of fraud can be divided into intra-organizational and extra-organizational categories:

**Intra-Organizational Fraud:** This category includes the cases of fraud committed by employees and managers inside an organization.

**Extra-Organizational Fraud:** This category includes the theft or abuse of organizational resources by individuals outside an organization.



**Figure 1.** Different forms of fraud committed by various fraudsters [8]

According to *Section 24* of Auditing Standards in Iran, financial statement fraud is classified as a form of intra-organizational fraud. This category includes cases of deception such as falsification of documents, manipulation or modification of accounting records or underlying

evidence for financial statements, incorrect presentation of events in financial statements, deliberate exclusion of events, misuse of accounting standards for measurement, identification, or classification, and intentional presentation or disclosure of financial information.

### 2.3. Data Mining

Hand *et al.* [9] defined data mining as the process of detecting and extracting knowledge from correct, novel, and incomprehensible patterns of big datasets. Emerging in the late 1980s, this method is now considered among ten knowledge development techniques that can integrate statistics, computer science, AI, machine learning, and visual representation of data [10]. It is widely used in medicine, engineering, finance, risk management, and fraud detection in particular.

***k*-Nearest Neighbors Algorithm:** This algorithm is considered among the simplest but the most important classification methods based on the idea of finding a specific number of nearest elements in a statistical population as the new element enters that population. The nearest datum to the new element in terms of different features should then be found and placed in the same category where the nearest elements exist. According to Yingquan *et al.* [11], the *k*-NN algorithm is a nonparametric method of classification for determining the distribution function in the distributed data. There is a training document or training datum for classification. This algorithm tries to find the similarity in the pre-classified training documents based on certain criteria. The classes of this algorithm will then be employed to predict the class of that training document by ranking the documents of each designated class [12]. Generally, the *k*-NN algorithm is a specific method of sample-based learning that deals with symbolic data. It is also considered a method of lazy learning that waits until a query is generalized beyond training data [13].

**Bayesian Network Algorithm:** The introduction of the Bayesian network dates back to the discovery of the Bayes formula in 1763 by an English priest named Thomas Bayes. According to the Bayes probability theorem, this algorithm estimates the probability of membership in a specific group [14].

The Bayes theorem is as follows:

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

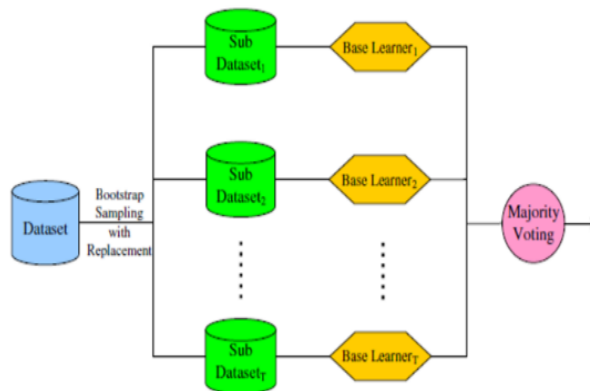
Where *X* and *Y* denote the observation (or a set of attributes) and the result (or the group label), respectively, to create a dataset. Moreover, *P*(*Y*|*X*) refers to the posterior probability of *X* at possible classes, whereas *P*(*Y*) represents the prior probability of each class without any information about *X*. Furthermore, *P*(*X*|*Y*) indicates the conditional probability of *X* with the probability of *Y*, whereas *P*(*X*) is basically the probability of observations.

To classify a new sample, *P*(*Y*|*X*) can be calculated for a specific group of *Y* to analyze which group has a greater value. The specific group of *Y* with the greatest value of *P*(*Y*|*X*) for a specific attribute of *X* is considered an estimate group for a new sample. Since *P*(*X*) yields the same results for any values of the specific group, it does not need to be calculated for any new samples; thus, it is considered constant [15].

**Support Vector Machine (SVM) Algorithm:** The SVM algorithm is a supervised learning classification method for solving classification or regression problems. Introduced by Vapnik (1995), this algorithm is based on the statistical learning theory and minimization of structural risks. It draws some hyperplanes in the space to optimally differentiate various data samples.

In other words, it distinguishes between the two groups in a way that they are the farthest from the nearest points from each group. The best hyperplane is the plane with the longest distance from both groups. This method classifies data by finding the best hyperplanes that distinguish all data of a group from data of the other group [16].

**Bagging Algorithm:** This algorithm is a collective learning method introduced by Breiman in 1996 for error reduction by employing a set of machine learning models of the same type. In the bagging algorithm, every classification method develops a model based on training data to detect differences of various classes. Instead of developing a model, this algorithm benefits from the models created by other classifiers and determines what class should be selected for the current sample by voting. Each class has access to the dataset. In this method, a subset of the main dataset is given to each classifier. In other words, each classifier monitors one part of the dataset (*i.e.*, features) to develop its model based on that accessible part of data—all features are not accessible to all classifiers [15].



**Figure 2.** Bagging Algorithm [17]

## 2.4. Grey Wolf Optimization Algorithm

The grey wolf optimization (GWO) is a metaheuristic algorithm inspired by the hierarchical structures and social behaviors of grey wolves while hunting. Following a simple process, this population-based algorithm can easily be generalized to large-scale problems. Grey wolves are considered apex hunters on top of the food pyramid.

This algorithm consists of three major phases:

- 1) Observing, tracking, and chasing a prey.
- 2) Approaching, surrounding, enclosing, and confusing the prey until it stops moving.
- 3) Attacking the prey [18].

## 2.5. Research Background

The reference reviewed by [19] created a fraud detection model utilizing the XGBoost algorithm, which aided in identifying fraud in a number of Middle Eastern and North African (MENA) companies. The sampling method algorithm (SMOTE) was employed to analyze the class imbalance issue in the dataset. To predict financial statement fraud, a variety of machine learning approaches were implemented in the Python programming language. Additionally, experimental results demonstrated that the XGBoost method outperformed the other algorithms in this study, including logistic regression (LR), decision tree (DT), and support vector machine (SVM), with an accuracy of 96.05%. The reference reviewed by [20] provided

a four-step artificial intelligence-based methodology for preventing corporate financial risk that would involve data preprocessing, feature selection, feature categorization, and parameter setting. Data for the financial index are gathered in the first stage, and pre-processing improves the quality of the designated data. In fact, the designated datasets are selected and optimized for features in the second stage, which builds a mathematical model through the chaotic grasshopper optimization algorithm (CGOA). The support vector machine then processes the classification of quantitative data through the condensed features. The SMA algorithm, which improves the SVM efficiency and accuracy, is the last step in the optimization process. The experimental findings demonstrated that, with an accuracy of 85.38%, the CGOA–SVM–SMA algorithm suggested in this study had superior prediction and decision-making capabilities as opposed to other models. The reference reviewed by [21] analyzed random forest, GBDT, XGBoost, and LightGBM machine learning models to create a financial statement fraud detection feature system for public businesses. They also developed an integrated feature selection technique for this purpose. The issue of unbalanced distribution was also resolved substantially, and the capacity to identify fraud was enhanced greatly by the addition of the SMOTE algorithm. GBDT had the best AUC performance and sensitivity among the four designated machine learning methods.

The reference reviewed by [22] extracted two nonfinancial ratios and 19 financial ratios by conducting a literature review, using snowball sampling, and interviewing experts. They then used an artificial neural network and a support vector machine for fraud risk prediction and detection. According to the results, the support vector machine outperformed the artificial neural network with the prediction power of 86%. The reference reviewed by [23] employed data preprocessing techniques in addition to feature selection of missing values, management of unbalanced classes, merged features, and distance correlation for feature selection with four classifiers of neural network, decision tree, extra trees, and random forest. They reduced 72 financial ratios to 18 ratios. According to the results, the 18 ratios selected by the features merged with the random forest classifier yielded an accuracy of 98.92%, which was higher than those of other methods. The reference reviewed by [24] used 41 financial and nonfinancial variables in a Bayesian network, a decision tree, a neural network, a support vector machine, and a combinatorial method for fraud risk detection. Their results indicated that the combinatorial method outperformed the other techniques with a prediction rate of 96.2% and a higher evaluation ability. The reference reviewed by [25] used five supervised methods, *i.e.*, feedforward multilayer neural network, probabilistic neural network, support vector machine, polynomial linear logarithmic model, and differential analysis with 18 financial data for fraud risk prediction in financial statements. Their results indicated that the feedforward multilayer neural network outperformed other methods in fraud risk detection with an accuracy above 90% in financial reports. The reference reviewed by [26] selected 23 financial ratios with available information in Iran to propose a novel approach to fraud risk detection in financial statements by searching empirical evidence. They then extracted 16 ratios as the best and most effective ratios by using the cross-entropy method. Moreover, they employed the logistic regression, genetic algorithm, and artificial bee colony (ABC) algorithm to classify companies as fraudulent and non-fraudulent categories. According to their results, the ABC algorithm outperformed the other methods in fraud risk prediction with an accuracy of 82.5%. The reference reviewed by [27] used a neural network and a support vector machine to extract appropriate variables for fraud prediction based on 22 financial and nonfinancial variables within an 11-year period. They obtained 10 and three variables from the neural network and the support vector machine, respectively. They also used four decision tree

techniques (*i.e.*, CHAID, CART, C5.0, and QUEST) to analyze the accuracy of fraud risk detection in financial statements. According to their results, 10 variables extracted by the artificial neural network and classification with the CART decision tree yielded the highest accuracy of fraud risk detection (90.21%) in financial statements. The reference reviewed by [28] predicted fraud risk in financial reports through logistic regression, support vector machine, multiple-criteria decision analysis, and artificial neural network. They utilized 10 financial ratios for fraud risk prediction. The results indicated that the artificial neural network outperformed the other methods in fraud risk prediction with an accuracy of 94.87%. The reference reviewed by [29] proposed a model for financial fraud risk prediction by conducting stepwise regression and elastic net tests through two steps in MATLAB. For this purpose, selected seven financial ratios were also used: ratio of working capital to asset, ratio of accounts receivable to sales, ratio of cash to current debt, ratio of inventory to current asset, ratio of debt to equity, ratio of gross income to asset, and absolute value of changes in current ratio. The logit test results indicated that 64.04% of the estimated model could be predicted. The reference reviewed by [30] used different data mining methods such as logistic regression, artificial neural network, and k-means clustering as well as various metaheuristic techniques such as distance-based and entropy-based ant colony algorithms and the genetic algorithm to detect cases of fraud risk through financial ratios. They tested each of the foregoing models at 82 Iranian companies. The results indicated that the distance-based ant colony algorithm outperformed the other methods. The reference reviewed by [31] analyzed the capabilities of six well-known statistical and machine learning models to detect financial statement fraud under the presumptions of misclassification costs and ratios of fraudulent to non-fraudulent organizations. The findings demonstrated that logistic regression and support vector machines outperformed artificial neural networks, bagging, and C4.5. Moreover, 6 out of 42 predictors (*i.e.*, auditor turnover, total optional accruals, four major international accounting firms, professional services, accounts receivable, meet or fail analyst forecasts, and unexpected employee productivity) were selected by classification algorithms. Hence, they can be used by experts to enhance fraud risk detection models. The reference reviewed by [32] analyzed how well data mining classification algorithms could be employed to spot businesses that produced false financial statements (FFS). It employed classification techniques such as decision trees, neural networks, and Bayesian networks to detect false financial statements. The Bayesian network model outperformed decision trees and neural networks in terms of classification accuracy, scoring 90.3% and 73.6%, respectively.

### **3. Research Hypothesis**

- 1) Feature reduction (*i.e.*, financial ratios) by the grey wolf optimization algorithm is more efficient in fraud risk detection than the lack of feature reduction.
- 2) The bagging algorithm is more effective than the other classifiers (*e.g.*, *k*-NN, Bayesian network, and support vector machine) in fraud risk prediction.

### **4. Research Methodology**

This is a descriptive-correlational study with a quantitative ex-post facto process. The statistical population included all companies listed in the Tehran Stock Exchange. The systematic conditional sampling method was employed to select the research sample. The fiscal years of companies were expected to end on March 20 (or March 21), and they were not selected from financial intermediaries such as investment companies, holdings, banks, and

insurance companies. The necessary data of research variables were expected to be available. Based on these conditions, 180 companies were selected.

Both Iranian and foreign studies (e.g., books and papers) were reviewed through notes in a library method to collect the necessary data regarding theoretical foundations and research background. The necessary data of variables were collected from financial statements and reports provided by independent auditors and authorized inspectors and published by the Tehran Stock Exchange. Rahavard Novin Software Suite and MS Excel were also utilized for essential calculations. Moreover, metaheuristic and data mining methods were used in MATLAB and DATALAB for data analysis and hypothesis testing.

## 5. Research Variables and Models

### 5.1. Dependent Variable

To define and detect fraud in financial statements as the dependent variable, *Audit Standard 240* entitled *Auditor's Responsibility* was reviewed along with the theoretical foundations of domestic and foreign studies regarding fraud risk detection to extract the most important cases of fraud:

- 1) Overestimating and underestimating incomes and assets
- 2) Overestimating and underestimating costs and debts
- 3) Restated financial statements and significant yearly moderations
- 4) Tax differences from tax areas and insufficiency of savings for performance tax
- 5) Stagnant assets and items such as inventory
- 6) The assumption of a company's nonstop activity for several consecutive periods is doubted, and an auditor's statement is conditional. However, the company is still supposed to present financial statements based on the continuity of its activities. For instance, consider a company in which production was stopped two years ago with no sales.
- 7) Misuse of accounting standards for identification, measurement, classification, presentation, and disclosure.

Some Iranian studies (e.g. [19, 23, 26, 27]) have confirmed the relationships between fraud cases and auditor statements. Hence, the paragraphs of condition and the other paragraphs of audit reports of companies with moderated statements (i.e., rejected statements, lack of statements, and conditional statements) were analysed thoroughly. Out of 1440 fiscal years (180 companies in 8 years), 532 fiscal years were identified as suspiciously fraudulent, whereas 908 fiscal years were identified as non-fraudulent. The suspiciously fraudulent companies were represented by 1, whereas the non-fraudulent companies were represented by 0.

### 5.2. Independent Variable

According to many Iranian and non-Iranian studies (e.g., [19, 20, 22, 23, 24]), financial ratios are capable of describing the importance of corporate features in relation to major events such as fraud. Financial ratios were used as the independent variables or fraud predictors of financial statements in this study. After a review of literature and theoretical foundations, financial ratios were extracted and classified as four categories of liquidity, leverage, efficiency, and profitability. In the initial analysis, some of the similar and inverted ratios were excluded. Finally, 96 financial ratios remained.



## 6. Hypothesis Analysis Methods

### 6.1. Evaluation Criteria and Measuring the Capabilities of the Proposed Pethods for Fraud Risk Detection in Financial Statements

The following evaluation criteria were employed to assess the proposed classifiers in fraud prediction:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

### 6.2. Analysing the Confusion Matrix of the Proposed Methods for Fraud Risk Prediction in Financial Statements

The quantities of rows and columns in a confusion matrix depend on the number of classes. There are two classes (*i.e.*, suspiciously fraudulent companies and non-fraudulent companies) in this study; hence, the confusion matrix includes the following elements:

True Positive (TP): This element indicates the suspiciously fraudulent financial statements identified correctly.

False Positive (FP): This element denotes the suspiciously fraudulent financial statements identified wrongly as non-fraudulent.

True Negative (TN): This element refers to the non-fraudulent financial statements identified correctly.

False Negative (FN): This element represents the non-fraudulent financial statements identified wrongly as suspiciously fraudulent.

### 6.3. The Receiver Operating Characteristic (ROC) of the Proposed Methods for Fraud Risk Detection in Financial Statements

The receiver operating characteristic (ROC) curve demonstrates the 2D presentation of results from the proposed methods. The  $x$ -axis and the  $y$ -axis represent the values of TP and FP, respectively. In this method, a common criterion is to calculate the area under curve of the ROC.

The efficiency of each algorithm was determined with the SVM classifier based on the values of accuracy, recall, precision, TP, and FP in the ROC.

## 7. Research Results

We now face big data with rapidly increasing features, the resultant information of which might be redundant, irrelevant, and obsolete [33]. The ratio of a sample size to the number of features should be appropriate in order to obtain reliable results for classification of fraudulent and non-fraudulent reports. Hence, feature selection is essential for complicated problems such as fraud detection. In addition, reducing redundant features can help retain a number of features including appropriate information, a process which usually improves learning, decrease computing costs, and enhance divisibility of the model [34].

The *k*-NN algorithm, Bayesian network, support vector machine, and bagging method were used as data mining algorithms in this study to classify companies as non-fraudulent and suspected of fraud once with all financial ratios and then with the financial ratios extracted by using the particle swarm optimization algorithm. The results were then saved as the tables extracted from a MATLAB simulator.

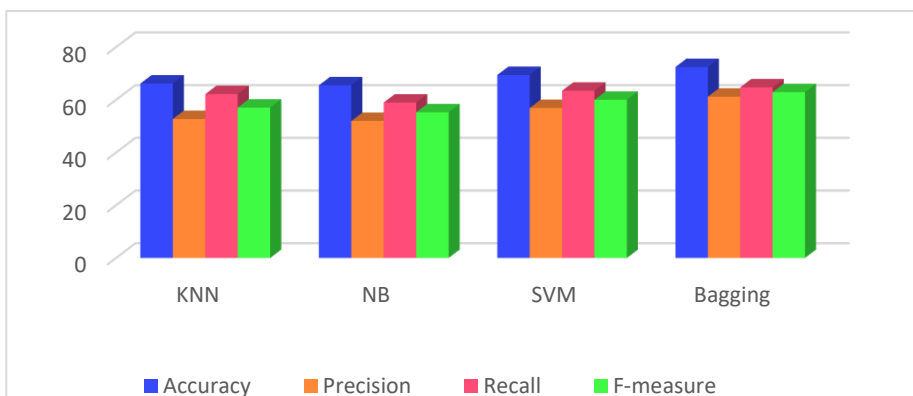
The learning techniques were first trained to analyze and evaluate the proposed algorithms. For this purpose, 70% of data (*i.e.*, 1008 data including 376 data of companies suspected of fraud and 632 data of non-fraudulent companies) were utilized as training data in MATLAB to calculate the training percentage of each model. Finally, the remaining 30% of data (*i.e.*, 432 data including 156 data of companies suspected of fraud and 276 data of non-fraudulent companies) were utilized as the test data in MATLAB to assess the algorithms and predict fraud risk.

### 7.1. Results of Evaluating Proposed Methods for Fraud Risk Detection without Feature Reduction

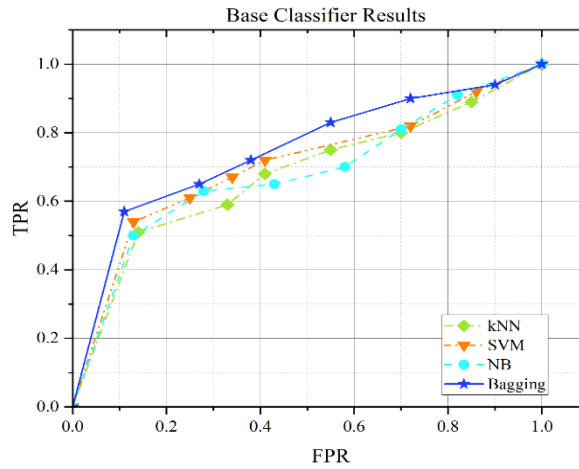
Table 1 and Figures 3–4 report the results of performance evaluation criteria, confusion matrix, and the ROC of classification methods with 96 financial ratios collected from 30 executions through test data.

**Table 1.** Results of performance evaluation, confusion matrix, and ROC of proposed methods with 96 financial classifications

Criterion	K-NN	Bayesian Network	SVM	Bagging
Accuracy	66.20%	65.51%	69.44%	72.45%
Precision	52.72%	51.98%	56.90%	61.21%
Recall	62.18%	58.97%	63.46%	64.74%
F-Measure	57.06%	55.26%	60%	62.93%
TP	97	92	99	101
TN	189	191	201	212
FP	87	85	75	64
FN	59	64	57	55
Efficiency (ROC)	67%	68%	71%	73.50%



**Figure 3.** Brief results of performance evaluating proposed method without feature reduction



**Figure 4.** Measuring and evaluating the efficiency of the all-financial ratios (*i.e.*, 96 financial ratios) with the proposed classification methods

The extracted values for the performance evaluation and the confusion matrix of the proposed classification methods are not relatively small and inappropriate with respect to all financial ratios. This finding indicates that appropriate features should be selected and utilized to classify companies and to improve results.

### 7.2. Selecting Financial Ratios through GWO

In the second step, the GWO algorithm (*i.e.*, a metaheuristic method) was employed in MATLAB to select the best financial ratios from 96 ratios. This algorithm is a metaheuristic optimization method inspired by the hierarchical structures and behaviors of grey wolves [18]. In this algorithm, each wolf is regarded as a solution to the problem in order to determine the best combination of financial ratios pertaining to non-fraudulent and fraud suspicious financial statements to accurately classify training samples and predict test samples. The solution that has the largest value with respect to the following fitness function will be considered the optimal solution.

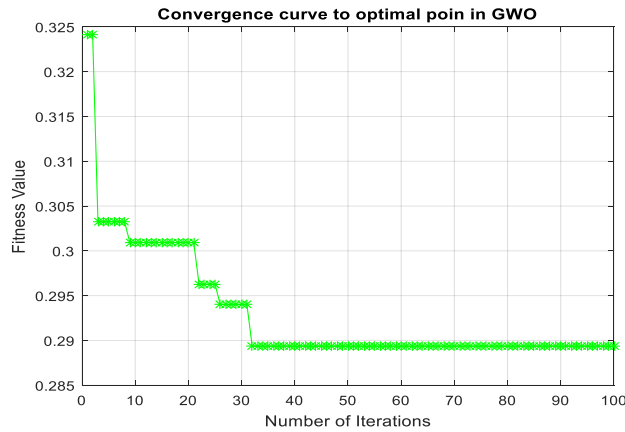
$$F = \text{Max } f(x) = \sum_{j=1}^M s_j x_j - p \times m \quad x_j \in \{0 \text{ and } 1\} \quad (6)$$

According to Table 2, nine financial ratios were selected as features based on the optimal solution in this algorithm:

**Table 2.** The financial ratios selected by the GWO algorithm

Financial Ratio	Iterations in 30 Executions	Financial Ratio	Iterations in 30 Executions
Total debts to total assets	18	Gross profit to total assets	22
Net profit to total assets	17	Cash balance to total assets	17
Working capital to total assets	21	Net profit to gross profit	18
Receivable accounts to sales	23	Accumulated profit and loss to equity	22
Current asset to current debt	15		

These financial ratios were selected as optimal features from the highly correlated features. The main criterion for evaluating the solution in the GWO is the detection error of financial statements; hence, the selected features will be more optimal if the fitness function value is smaller. Figure 5 demonstrates the convergence of the fitness function values on the optimum by the GWO algorithm.



**Figure 5.** The convergence of the fitness function on the optimum in the GWO algorithm

According to Figure 4, the fitness function values of the GWO algorithm in the feature subset selection problem converged on the optimum with an error rate of zero as the iterations increased. After 100 iterations, the fitness value of this algorithm was obtained 0.2940, and the accuracy of financial ratios selected by the GWO algorithm was % 70/60 for training data to detect non-fraudulent financial statements and the statements suspected of fraud. This algorithm yielded the best financial ratios after 31 iterations at a high speed.

### 7.3. Validity of GWO

The test data were employed to analyze the validity of financial ratios extracted by the GWO algorithm. Table 3 reports the results.

**Table 3.** The validity of the GWO algorithm

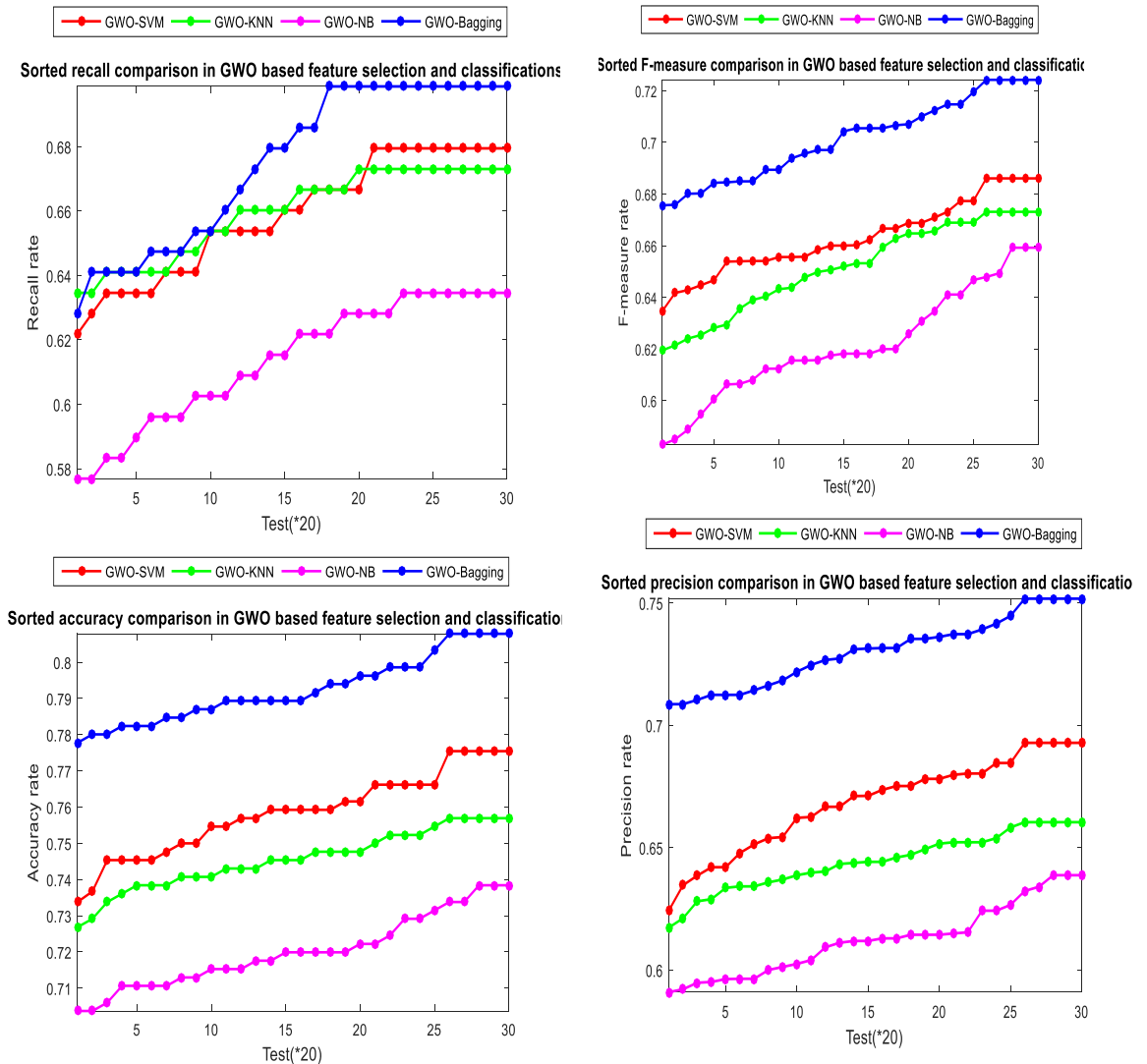
Detection Result	Non-Fraudulent	Suspected of Fraud	Total	Precision
Non-fraudulent	240	36	276	78.41%
Suspected of Fraud	47	109	156	

### 7.4. Results of Evaluating the Proposed Methods and the Confusion Matrix for Fraud Detection through Financial Ratios Selected by Grey Wolf Optimizer

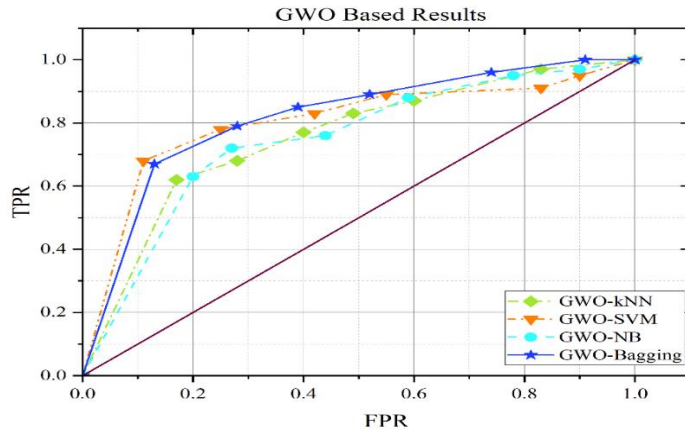
Table 4 and Figures 6–7 report the results of performance evaluation criteria, confusion matrix, and ROC of classification methods with the financial ratios collected from the grey wolf optimization algorithm in 30 iterations through test data.

**Table 4.** The results of performance evaluation, confusion matrix, and ROC of proposed methods with financial ratios extracted from the GWO algorithm

Criterion	K-NN	Bayesian Network	SVM	Bagging
Accuracy	74.55%	72.01%	75.83%	79.25%
Precision	64.43%	61.23%	66.80%	73.01%
Recall	65.96%	61.37%	65.88%	67.52%
F-Measure	65.15%	62.17%	66.31%	70.13%
TP	105	99	106	109
TN	222	220	229	240
FP	54	56	47	36
FN	51	57	50	47
Efficiency (ROC)	75.60%	74.60%		80.30%
				81.70%



**Figure 6.** Comparison of methods in financial ratios extracted by GWO



**Figure 7.** Measuring and evaluating the efficiency of financial ratios extracted by the GWO with the proposed classification methods

### 8. Conclusion and Suggestions

Fraud is now committed with complicated and organized schemes; therefore, many fraud cases are costly and lead to risks and mistakes in decisions made by investors, creditors, and other users. They also cause serious non-financial effects, especially the loss of accounting credibility. Many of these fraud cases are left undetected. Therefore, it is essential to develop efficient methods of detecting fraud in financial statements. There are not many specific independent variables affecting fraud prediction in financial statements. Since the analysis of numerous variables can be time-consuming and redundant, they will cause confusion and error in fraud detection.

The grey wolf optimization algorithm was employed in this study to reduce and extract appropriate financial ratios for fraud detection in financial statements. Statistical methods are more capable of prediction in linear continuous data than in nonlinear discrete data [35]. These methods are also less likely to succeed in fraud risk detection. Described by [36] as one of the top ten technologies and the process of discovering unknown relationships and patterns inside data, data mining methods were employed in this study for fraud risk detection through the extracted financial ratios [37]. According to the brief results of performance evaluation criteria for financial ratios and the outputs of the proposed algorithms in the analysis of hypotheses, the first research hypothesis was confirmed. In other words, the reduction of features (*i.e.*, financial ratios) is more efficient in fraud risk detection than the lack of feature reduction. Table 5 indicates that the financial ratios extracted by the grey wolf optimization algorithm are consistent with the findings reported by previous studies.

**Table 5.** The results of analyzing the extracted financial ratios in comparison with the previous ratios

Financial Ratio	Previous Studies
Total debts to total assets	[19, 20, 21, 24, 25, 27]
Net profit to total assets	[21, 27]
Working capital to total assets	[23, 25, 26]
Receivable accounts to sales	[19, 23, 25]
Current asset to current debt	[21]

Gross profit e to total assets	[25, 26]
Cash balance to current debt	[19, 20, 23]
Net profit to gross profit	[23]
Accumulated profit and loss to equity	[27]

According to the results of evaluating the proposed methods, the superiority of GWO–bagging was confirmed by performance evaluation, confusion matrix, and ROC. Hence, the second research hypothesis was confirmed. In other words, bagging classification algorithms outperformed  $k$ -NN, Bayesian network, and support vector machine in fraud risk detection. Many of the previous studies (e.g., [19 ,20, 22, 24, 21, 23, 27]) confirmed the superiority of data mining methods to other techniques in fraud risk detection.

### 8.1. Suggestions

The identification of fraudulent companies that manipulate financial ratios is a challenging, specialized, and time-consuming task. Moreover, the majority of financial information users lack the essential and sufficient expertise for this purpose. Hence, an organization or an institution should be established to address fraudulent financial reporting more seriously than ever before. In addition, a specialized association should also be formed to identify fraudulent companies and disclose their information publicly. Furthermore, the esteemed legislative organizations and institutions should revise trade laws, devise controlling mechanisms and legally binding frameworks, adopt preventive and punitive measures, and increase penalties to reduce the fraud risk in financial statements.

## References

- [1] Rastatter, S., Moe, T., Gangopadhyay, A., & Weaver, A. Abnormal traffic pattern detection in real-time financial transactions. No. 827. EasyChair, (2019) 1-7.
- [2] Occupational Fraud. A Report to Nations, <https://acfepublic.s3.us-west-amazonaws.com>. (2016, 2018, 2020, 2022).
- [3] Chimonaki, C., Papadakis, S., Vergos, K., & Shahgholian, A. Identification of financial statement fraud in Greece by using computational intelligence techniques. *Enterprise Applications, Markets and Services in the Finance Industry. FinanceCom 2018. Lecture Notes in Business Information Processing* 345 (2019) 39-51.
- [4] Etemadi, H., & Zolqhy, H. Using logistic regression to identify fraudulent financial reporting. *Journal of Audit Science* 13 (2013) 163-145.
- [5] Sadgali, I., Sael, N., & Benabbou, F. Performance of machine learning techniques in the detection of financial frauds. *Procedia computer science* 148 (2019) 45-54.
- [6] Yao, J., Pan, Y., Yang, S., Chen, Y., & Li, Y. Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: a multi-analytic approach. *Sustainability* 11 (2019) 1579.
- [7] Auditing Standards Committee. *Principles and Regulations of Accounting and Auditing: Auditing Standards, Audit Organization Publications*, Tehran, Iran. (2015).
- [8] Goldmann, P. D., & Kaufman, H. *Anti-Fraud Risk and Control*. USA: Hoboken. (2009).
- [9] Han, J., Kamber, M., & Pei, J. *Data mining concepts and techniques* third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University (2012).
- [10] Rahnamay Roodposhti, F. Data mining & financial fraud. *Journal of Management Accounting and Auditing Knowledge* 1 (2012) 17-34.
- [11] Wu, Y., Ianakiev, K., & Govindaraju, V. Improved k-nearest neighbor classification. *Pattern recognition* 35 (2002): 2311-2318.
- [12] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. KNN model-based approach in classification. *On The Move to Meaningful Internet Systems: OTM Confederated International Conferences*, (2003) 986-96.
- [13] Kuncheva, L. I. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, (2014).



- [14] Leung, K. M. Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering (2007) 123-156.
- [15] Shinde, A., Sahu, A., Apley, D., & Runger, G. Preimages for variation patterns from kernel PCA and bagging. *Iie Transactions* 46 (2014) 429-456.
- [16] Pradhan, A. Support vector machine-a survey. *International Journal of Emerging Technology and Advanced Engineering* 2 (2012): 82-85.
- [17] Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. Sentiment classification: The contribution of ensemble learning. *Decision support systems* 57 (2014) 77-93.
- [18] Mirjalili, S., & Lewis, A. S-shaped versus V-shaped transfer functions for binary particle swarm optimization. *Swarm and Evolutionary Computation* 9 (2013) 1-14.
- [19] Ali, A. A., Khedr, A. M., El-Bannany, M., & Kanakkayil, S. (2023). A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. *Applied Sciences*, 13(4), 2272.
- [20] Lei, Y., Qiaoming, H., & Tong, Z. (2023). Research on Supply Chain Financial Risk Prevention Based on Machine Learning. *Computational Intelligence and Neuroscience*.
- [21] Chen, Y. (2023). Financial Statement Fraud Detection based on Integrated Feature Selection and Imbalance Learning. *Frontiers in Business, Economics and Management*, 8(3), 46-48.
- [22] Kamrani,H.,& Abedini, B. Formulation of Financial Statement Fraud Detection Model Using Artificial Neural Network and Support Vector Machine Approaches in Companies Listed in Tehran Bahador Stock Exchange. *Journal of Management Accounting and Auditing Knowledge* 11 (2022) 285-314.
- [23] Cheng, C. H., Y. F. Kao & H. P. Lin. A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. *Applied Soft Computing*108 (2021) 107487.
- [24] Rezaei, M., Nazemi Ardakani, M., & Naser Sadr Abadi, A. (2020). Fraud Detection in Financial Statements through Audit Reports of Financial Statements. *Management Accounting Journal*, 13 (2020) 141–153.
- [25] Omid, M., Q. Min., V. Moradinaftchali & M. Piri. (2019). The efficacy of predictive methods in financial statement fraud. *Discrete Dynamics in Nature and Society*, 2019 (2019) 1-12.

- [26] Tashdidi, E., Sepasi, S., Etemadi, H., & Azar, A. Proposing a Novel Approach to Fraud Prediction and Detection in Financial Statements through Bees Algorithm. *Journal of Accounting Knowledge*, 10 (2019) 139–167.
- [27] Jan, C. L. An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. *Sustainability*, 10 (2018) 513.
- [28] Omar, N., Z. A. Johari & M. Smith. (2017). Predicting fraudulent financial reporting using artificial neural network. *Journal of Financial Crime*, 24(2):362-387.
- [29] Zareh Bahmanmiri, M., & Malekian Kaleh Basi, E. (2015). Fraud Prediction in Financial Statements through Financial Ratios. *Financial Management Outlook*, (12): (2017) 362-87.
- [30] Kazemi, T. Identifying Cases of Fraud Risk in Financial Statements of Iran and Evaluating Fraud Risk Detection Methods. *Doctoral Dissertation Doctoral*. (2016).
- [31] Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory*, 30(2), 19-50.
- [32] Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, 32(4), 995-1003.
- [33] Khalid, S., T. Khalil, and S. Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*. IEEE, (2014).
- [34] Vieira, S. M., Sousa, J. M., & Runkler, T. A. (2010). Two cooperative ant colonies for feature selection using fuzzy models. *Expert Systems with Applications*, 37(2010), 2714-23.
- [35] Ranganathan, P., C. S. Pramesh & R. Aggarwal. Common pitfalls in statistical analysis: Logistic regression. *Perspectives in clinical research* 8 (2017) 148.
- [36] Larose, D. T. *An introduction to data mining*. Traduction et adaptation de Thierry Vallaud. (2005).
- [37] Berry, M. J., & Linoff, G. S. *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, (2004).