Fraud Prediction in Financial Statements through Comparative Analysis of Data Mining Methods

Zahra Nemati
1Department of Accounting, Zanjan Branch, Islamic Azad University, Zanjan, Iran.
znemati5879@gmail.com

Ali Mohammadi
Department of Accounting, Zanjan Branch, Islamic Azad University, Zanjan, Iran.
(Corresponding Author)
ali_mohammadi93@yahoo.com

Ali Bayat
Department of Accounting, Zanjan Branch, Islamic Azad University, Zanjan, Iran.
ali.bayat22@yahoo.com

Abbas Mirzaei
Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran.
a.mirzaei@iauardabil.ac.ir

Submit: 07/02/2023 **Accept:** 30/09/2023

ABSTRACT

Fraud increases business risks and costs, creates investor distrust, and questions the professional competence and credibility of accounting. Hence, this study aims to employ data mining methods for fraud risk prediction at the companies listed in the Tehran Stock Exchange within the 2014–21 period. For this purpose, 96 financial ratios were collected by reviewing theoretical foundations and research literature. The proposed classifiers such as the *k*-nearest Neighbors algorithm, Bayesian network, support vector machine, and bagging classifier were adopted for fraud prediction. The performance of all classifiers were evaluated relatively poor. Therefore, financial ratios were reduced to enhance the proposed classifiers through the particle swarm optimization algorithm. In fact, 11 effective financial ratios were extracted with a precision of 72.92% and a prediction accuracy validity of 84.82 %. The extracted ratios were then reevaluated by the proposed classifiers for fraud prediction. According to the reevaluation results, all of the proposed methods improved with the extracted financial ratios. The research results indicated that the bagging classifier yielded the highest precision and accuracy, *i.e.*, 84.28% and 76.85%, respectively, and the lowest prediction error, *i.e.*, 23.15%. It was also 87% efficient in fraud prediction.

Keywords: *k*-nearest neighbors algorithm, Bayesian network, Support vector machine, Bagging classifier, Particle swarm optimization algorithm

1. Introduction

Fraud has many adverse outcomes in economic, cultural, and social aspects. In fact, corruption and fraud denote the abuse of power to gain personal or group advantages due to the lack of control over power, distortion of power, and the lack of executive guarantees. Moreover, corruption and fraud are known as unusual crimes that target the moral values and culture of society as well as the social policies of a state, thereby disrupting national competition and economic growth. They can also negatively affect business relations and neutralize the attempt at mitigating poverty and social discrimination (Umar & Purba, 2020).

There is an alarming rise in the number of reports published on financial fraud worldwide. Since 2009, these reports have substantially proliferated. According to the 2018 global economic crime and fraud survey, 49% of organizations were victims of economic crimes worldwide in 2016 and 2017 (Rastatter *et al.*, 2019).

Financial statements are analytical reports published periodically by financial institutions that explain their performance from various angles. Since these reports are the most important decision-making tools for many stakeholders, creditors, investors, and even auditors, some institutions may deceive people and commit fraud by manipulating these documents. Financial statement fraud detection aims to detect anomalies caused by these distortions and to distinguish suspected fraudulent reports from non-fraudulent ones (Aftabi *et al.*, 2023).

Although fraud in financial statements accounts for only 9% of cases of crime in reports, its average damage rate is $593,000 per fraud, a figure which is the costliest case of financial crimes (Association Chief Police Officers, 2022).

Prevention, detection, and investigation of fraud in financial statements of companies have now become new concerns of accounting more than anything else in the world. Nearly all organizations have somehow encountered different cases of fraud ranging from a negligible theft by an employee to fraudulent financial reporting. Major fraud in financial statements can have considerably adverse effects on the market value of a business and its credibility and ability to achieve strategic goals, leading finally to bankruptcy and loss of tens of thousands of job opportunities. In society, it can also reduce the financial market efficiency, destroy the public trust in accounting and auditing, and decline economic developments (Chimonaki *et al.*, 2018).

Since 2004, Transparency International Organization has reported financial corruption in different countries on a yearly basis. Countries are ranked in financial corruption based on scores ranging from 1 (*i.e.*, the highest level of corruption) to 100 (*i.e.*, no corruption). According to the evaluation standards by this organization, a score below 50 indicates a corrupt country. Iran was ranked 25th by this organization in 2021. Regarding financial corruption and its expansion, Iran has been ranked 150th out of 180 countries. Compared with the statistics published by this organization in 2017, Iran has descended 20 ranks, a decline which indicates the substantial growth of corruption in this country.

In most of the developed countries, there is an official organization, *e.g.*, Association of Certified Fraud Examiners in the US, to report statistics on the emergence of fraud and to introduce fraudulent companies. However, despite the importance of fraud in financial statements, there is no legal institution in Iran to directly investigate and discover fraud or to report the list of fraudulent companies for detection of fraud cases in financial statements. With advances in technology and high-speed communication networks, methods of fraud have now become so complicated that it is now easier to commit fraud but more difficult to detect it. In fact, fraudsters now act intelligently and quickly (Sadgali, Saela & Benabbo, 2019). Hence, it is a very difficult and complicated but important task to detect fraud. Thus, studies have gradually started using artificial intelligence techniques rather than conventional methods and statistical analysis due to their reliance on restrictive hypotheses such as normal distribution and high classification error rates (Yao *et al.*, 2019). Given the importance of fraud and its correct prediction in financial statements, this study aims to find the best method for detecting fraudulent companies by using useful and effective financial ratios through data mining methods which are classified as artificial intelligence techniques.

Theoretical Foundations and Hypotheses

Definition of Fraud

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

Audit standards present a specific definition of fraud. According to Section 24 of Audit Standards, distortion of financial statements can ensue from fraud or mistake. In the definition provided by this section of Audit Standards, fraud refers to any intentional or deceiving action taken by one or several managers, employees, or third parties with the purpose of gaining an unfair or illegal advantage. Although fraud has a pervasively legal concept, what concerns an auditor includes the fraudulent actions that lead to substantial distortion of financial statements (International Accounting Standards Committee, 2020).

Fraud has a wide range of legal implications. In general, however, it is an intentional act committed by an individual or a group to gain unfair and illegal benefits. Furthermore, violation is the misconduct that refer to the infringement of laws, regulations, and organizational procedures as well as disregard for market and business ethics (Hosseini, 2021).

According to Ibadin and Dikemor (2020), fraud has an extensive concept that refers to the acquisition of illegal benefits through deliberate deception and has different forms such as financial corruption, fraudulent reporting, and abuse of assets.

Fraud Detection Methods

Every deceiving activity usually starts with minor cases; however, if left undetected, it will escalate to major cases. Due to the growing number of deceiving events, it is essential to detect fraud in the first place. Organizations have made many efforts to detect fraud in recent years. The Association of Certified Fraud Examiners (ACFE,2022) in the US introduces fraud detection methods, the usage of each method, and the costs of each method every two years (Figures 1 and 2). The highest percentage of using detection methods belongs to the acquisition of confidential information (*i.e.*, informing), which can be disseminated by employees, sellers, and buyers. This method is not as costly as some other methods such as dissemination of information by legal authorities and external auditors. Internal audit, managerial investigations, random checking, and calculation of differences in accounts come at the next ranks.

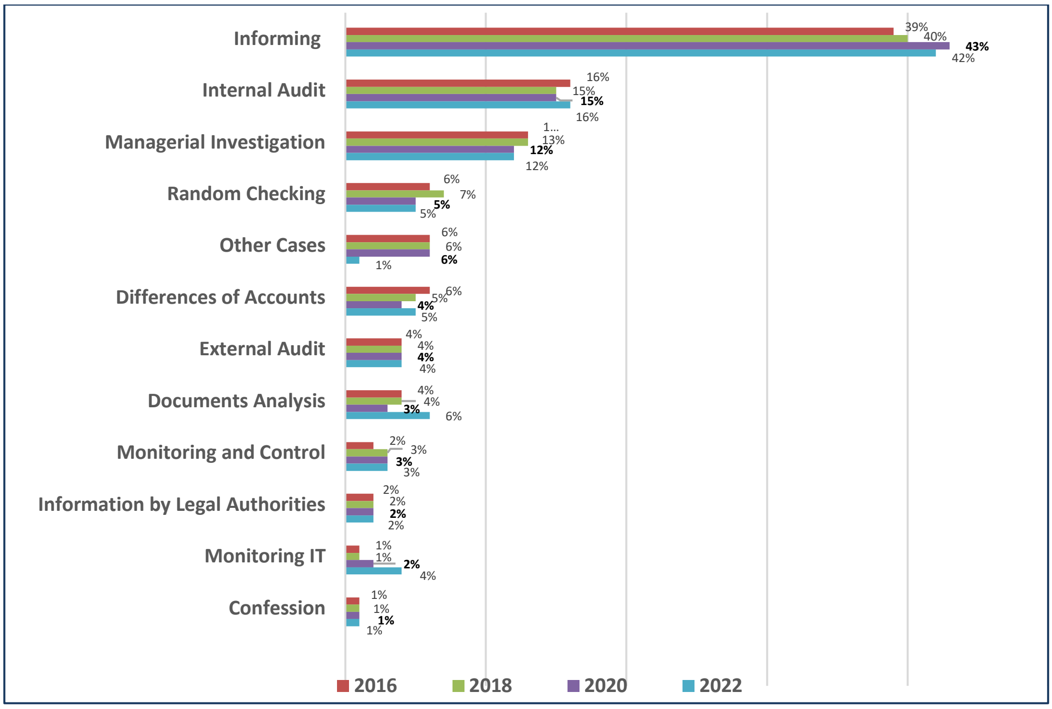

Figure 1. Fraud Detection Methods (ACFE,2022)

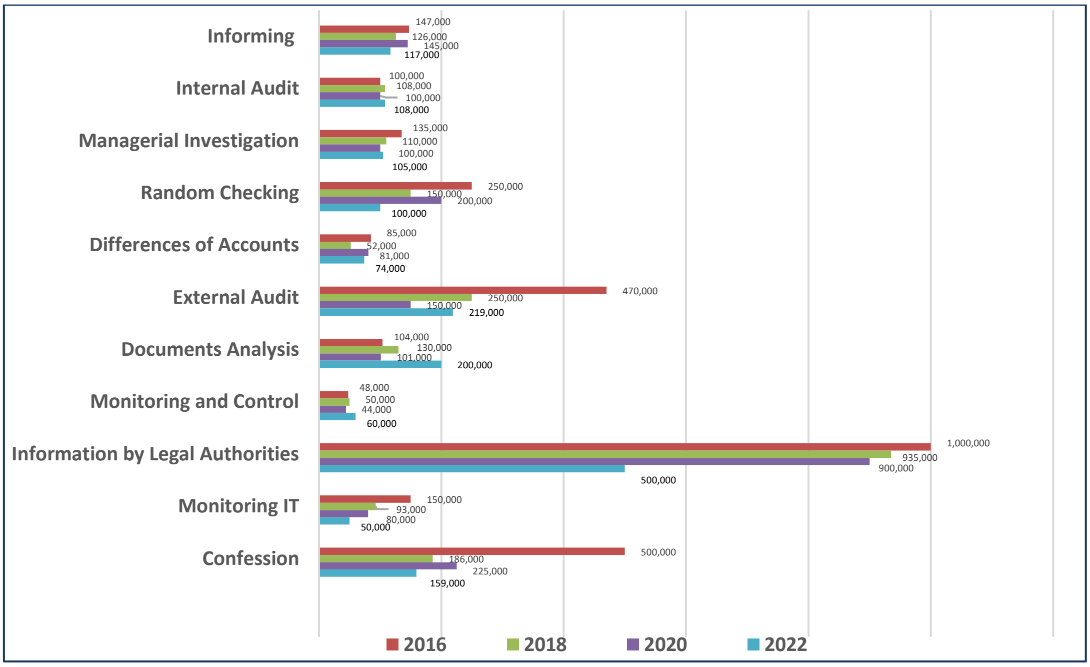

Figure 2. Costs of Fraud Detection Methods (ACFE,2022)

Data Mining

Hand *et al.* (2006) defined data mining as the process of detecting and extracting knowledge from correct, novel, and incomprehensible patterns of a big dataset. This method emerged in the late 1980s and is now known as one of the ten forms of developing knowledge that integrates statistics, compute science, AI, machine learning, and visual representation of data (Rahnama Roudposhti, 2012). It is widely used in medicine, engineering, finances, risk management, and especially fraud detection.

***k*-Nearest Neighbors Algorithms:** This algorithm is one of the simplest but the most important methods of classification based on the idea of finding a specific number of nearest elements in a statistical population as the new element entering 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 near elements exist. According to Yingquan *et al.* (2002), this algorithm is a kind of nonparametric classification for obtaining the distribution function from the distributed data. There is a training document or training data for classification, and this algorithm find the similarity among the pre-classified training documents based on a criterion. The classes of this algorithm will then be employed to predict the class of that training document by scoring the documents of each designated class (Guo *et al.*, 2002). Generally, the *k*-nearest neighbors algorithm is a specific case of sample-based learning that deals with symbolic data. This method is also a case of lazy learning that waits until a query is generalized beyond training data (Kuncheva, 2014).

Bayesian Network Algorithm: The Bayesian network dates back to the discovery of the Bayes formula in 1763 by an English priest named Thomas Bayes. Based on the Bayes probability theorem, this algorithm estimates the probability of membership in a specific group (Leung, 2007).

The Bayes theorem is as follows:

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

Where *X* and *Y* denote the observation (or a set of attributes) and the result (or the group label), respectively, to yield a dataset. Moreover, *P(Y|X)* indicates the posterior probability of *X* at possible classes, whereas *P(Y)* denotes the prior probability of each class without any information about *X*. Furthermore, *P(X|Y)* refers to 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 result for any values of the specific group, it does not need to be calculated for any new sample; thus, it is considered constant (Shinde *et al.*, 2014).

Support Vector Machine Algorithm: The SVM algorithm is a supervised learning classification method that can be employed to solve 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 distinguish between different data samples. In other words, it differentiates 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 (Pradhan, 2012).

Bagging Algorithm: This algorithm is a collective learning technique that was introduced by Breiman in 1996 for error reduction by using a set of machine learning models of the same type. In the bagging algorithm, every classification method develops a model on training data to perceive differences of various classes. Instead of developing a model, this algorithm benefits from the models created by the 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 observes one part of the dataset (*i.e.*, features) to develop its model based on that accessible part of data—all features are not given to all classifiers (Shinde *et al.*, 2014).

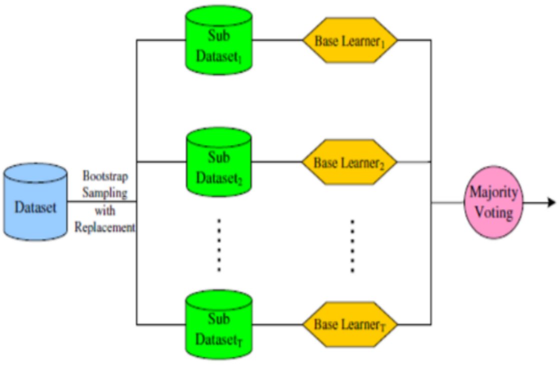

Figure 3. Bagging Algorithm (Wang *et al.*, 2014)

Particle Swarm Optimization Algorithm

In the early 1990, many studies were conducted on the social behaviors of animal groups. Inspired by those studies, Eberhart and Kennedy (1995) introduced the particle swarm optimization (PSO) algorithm, which is an appropriate method of optimizing nonlinear continuous functions. This algorithm is adapted from the simulated flying behavior of a flock of birds. It can be employed to solve a wide variety of optimization problems (El-Shorbagy & Hassanien, 2018). A flock of birds hunt for food randomly in a space where there is only one piece of food. None of the birds know where the food is. Following the bird that is nearest to the food can be an efficient strategy, which is the main theme of this algorithm. Called a particle, every solution in this algorithm represents a bird in the flock of flying birds. It has a fitness value calculated by a fitness function. It also has a velocity that is responsible for directing the particle. Each particle keeps moving in the problem space by following the optimal particles in the current state. In fact, a group of particles first emerge randomly and then try to find the optimal solution by updating the generations (Nath, Mishra, Kar, Chakraborty & Dey, 2014).

Financial ratios are expected to be effective in fraud risk prediction in financial ratios through novel methods. Hence, the following hypotheses were developed:

1) Reducing features (*i.e.*, financial ratios) can be more effective than not reducing features in fraud risk prediction.
2) The bagging algorithms is more effective than the other classifiers (*e.g.*, *k*-nearest neighbors, Bayesian network, and support vector machine) in fraud risk prediction.

Research Background

Ali *et al.* (2023) 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%.

Lei *et al.* (2023) 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.

Chen (2023) 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.

Aftabi *et al.* (2023) proposed a novel approach based on the generative adversarial networks (GAN) framework, which compiles a new dataset from the annual financial statements of ten Iranian banks and extracts three types of features. According to experimental results, the proposed method outperformed other classification techniques in generating synthetic suspected fraud samples with extreme gradient boosting (XGBoost) (99% accuracy) and SVM (100% accuracy). Furthermore, comparative performance with supervised models was more effective in accurately detecting suspected fraud samples than with unsupervised models.

Xiuguo and Shengyong (2022) proposed a financial fraud detection system by using the deep learning model based on a combination of numerical features and textual data. They utilized financial statements and managerial reports to extract financial indices and textual features, respectively, in management discussion and analysis (MD&A) through a lexical vector of annual reports on 5130 Chinese companies. They then employed deep learning models and compared their outputs with numerical data, textual data, and composite data. According to the results, the proposed method with the LSTM classifier and the GRU classifier yielded precisions of 94.98% and 94.48%, respectively, and outperformed the conventional classifiers.

Kamrani and Abedini (2022) extracted two nonfinancial ratios and 19 financial ratios by reviewing the research literature, using the snowball sampling method, and interviewing experts. They predicted and detect fraud risk through a neural network and a support vector machine. The results indicated that the support vector machine yielded a precision of 86% and outperformed the neural network.

Cheng *et al.* (2021) first used data preprocessing methods for feature selection to predict fraud risk in financial ratios. They then employed three feature selection methods of missing values, unbalanced class management, and merged features to reduce 72 financial ratios to 18 financial ratios. After that, they used four classifiers (*i.e.*, neural network, decision tree, additional trees, and random forest) to classify companies as fraudulent and non-fraudulent categories. According to their results, 18 financial ratios selected through the merged feature integrated with the random forest classifier had a precision of 98.92%, which was higher than those of other methods.

Gupta and Mehta (2021) used machine learning techniques and statistical methods for fraud detection in financial statements. Their results indicated that logistic regression, probit regression, neural network, decision tree, SVM, and fuzzy method yielded precisions of 71.5%, 89.5%, 71.7%, 73.6%, 90.4%, and 86.8%, respectively. Therefore, machine learning approaches outperformed statistical methods in fraud risk prediction at companies, especially when insufficient data are accessible to the sample.

Youkhneh Alghiani *et al.* (2021) integrated classic data mining, ANFIS, and metaheuristic algorithm to predict fraud risk in financial tax reports. The results indicated that using different optimization algorithms in the data mining approach increased the prediction power of the financial tax reporting detection model. In fact, the particle swarm optimization algorithm led to the most optimal model.

Rezaie et al. (2021) employed the CRISP approach to predict the financial statement fraud risk at the companies listed in the Tehran Stock Exchange. They reported that 40 independent financial and non-financial variables affected fraud. Four artificial intelligence techniques, e.g., decision tree, neural network, support vector machine (SVM), and AdaBoost–SVM, were adopted for fraud risk prediction. Finally, the CRISP approach was more effective than other techniques in predicting financial statement fraud risk with 82% accuracy.

Rezaei et al. (2020) used 41 financial and non-financial variables for fraud detection by using a Bayesian network, a decision tree, a neural network, a support vector machine, and a hybrid method. According to their results, the hybrid method outperformed the other techniques in precision and evaluation with a prediction rate of 96.2%.

Hidayattullah *et al.* (2020) used machine learning based on metaheuristic optimization for fraud detection in financial statements of Indonesian companies. For this purpose, they first selected 18 financial ratios with available information. Using the principal component analysis, they then extracted 10 financial ratios utilized in classification. After that, they employed several machine learning approaches based on metaheuristic optimization to develop models for fraud detection in financial statements. They used the genetic algorithm to reduce financial variables and employed the support vector machine and the optimized backpropagation neural network (BPNN) for classification. The financial ratios extracted by the genetic algorithm with a vector machine classifier yielded a precision of 96.15% and outperformed the other methods.

In a study entitled *Performance of Machine Learning Models in Fraud Detection*, Sadgali *et al.* (2019) analyzed data mining methods for fraud detection. According to their results, the probabilistic neural network (PNN) yielded the highest precision (98.09%).

Tashdidi *et al.* (2019) reviewed the empirical evidence and selected 23 financial ratios with available information in Iran to propose a novel approach to fraud detection in financial ratios. They then employed the cross entropy method to extract 16 ratios as the best and most effective ratios. They used logistic regression, genetic algorithm, and bees algorithm to classify companies as fraudulent and non-fraudulent categories. According to their results, the bees algorithm outperformed the other methods in fraud prediction with a precision of 82.5%.

Yao *et al.* (2018) proposed a model for fraud detection in financial statements through data mining methods. They reduced 17 extracted financial ratios to 6 and 5 financial ratios by using the PCA and Xgboost, respectively. They then employed a support vector machine classifier, a random forest, a decision tree, an artificial neural network, and a logistic regression to classify companies as fraudulent and non-fraudulent categories through the extracted financial ratios. According to their results, the support vector machine and the random forest yielded the best precision (71.67%) and the worst precision (68.17%), respectively.

Ebrahimi and Khajavi (2017) adopted the correlation-based feature selection method to select the variables with the greatest effects on fraud detection in financial ratios. For this purpose, they used 40 financial and non-financial ratios. The results indicated the usefulness of cash ratios, interest cover, accounts payable to total assets, inventory to net sales, sales logarithm, net profit to sales, and current assets to total assets. They also employed data mining methods such as artificial neural network, Bayesian network, and random forest for fraud prediction. The results indicated that the random forest algorithm outperformed the other techniques with a precision of 96.77%.

Kazemi *et al.* (2016) used different data mining methods such as logistic regression, artificial neural network, and *k*-mean 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 in 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.

Research Methodology

In this applied descriptive-correlational quantitative ex-post facto study, the necessary information was obtained from the review of theoretical foundations and both domestic and foreign research literature including books and papers through the desk method. Financial statements, reports published by independent auditors and legal inspectors of Iran Securities and Exchange Organization, and Rahavard Novin Software Suite were employed to collect the necessary data. Calculations were then performed in Excel. Moreover, metaheuristic and data mining methods were employed to analyze data and test research hypotheses in MATLAB and DATALAB.

Statistical Population and Sample

The statistical population included the companies listed in the Tehran Stock Exchange within 2014–2021 period. The companies meeting all of the following inclusion criteria were selected as the research sample:

1) The companies should be listed in the Tehran Stock Exchange until March 20, 2014, and their names should not be excluded from the Tehran Stock Exchange during the research period.
2) Their fiscal years should end on March 20 without any change within the research period.
3) They should not be among financial intermediaries such as investment companies, holdings, banks, and insurances.
4) The necessary information including financial statements and independent auditing reports should be available for the research period.

Given the inclusion criteria, 180 companies were selected.

Research Variables and Measurement Methods

Dependent Variable: Fraud in financial statements was defined as the dependent variable by analyzing the Audit Standard 240 known as the auditor's responsibility and reviewing the theoretical foundations of domestic and foreign studies on fraud. The cases of fraud were then identified, and the most important cases were extracted and listed as below:

1) Overrepresentation and underrepresentation of incomes, costs, assets, and debts
2) Repeated financial statements and substantially annual moderations
3) Tax discrepancies in taxation areas and insufficiency of reserves for performance tax
4) Stagnant assets and articles such as inventory
5) The assumption of continuous activity is questioned in a company for several consecutive periods, and the auditor's statement is conditional; however, the company still provides its financial statements based on the continuity of activities. For instance, consider a company which stopped operating two years ago without having any sales.
6) The misuse of accounting standards regarding identification, measurement, classification, presentation, and disclosure

Some studies have confirmed the relationship between fraud and auditor statement. Hence, according to the foregoing cases of fraud, paragraphs of conditions and other paragraphs of auditing reports of companies with moderated statements (*i.e.*, rejected statement, no statement, and conditional statement) were analyzed comprehensively. Finally, 532 out of 1440 year-companies (180 companies in 8 years) were identified as suspected of fraud, whereas 908 year-companies were identified as non-fraudulent. The companies suspected of fraud were represented by 1, and the non-fraudulent companies were represented by 0.

Independent Variable: Financial ratios were used as independent variables or predictor variables of fraud in financial statements. After the theoretical foundations and the research literature were reviewed, financial ratios were extracted and classified as four categories called liquidity, leverage, efficiency, and profitability. Some of the similar and inverted ratios were excluded in the primary analysis. Finally, 96 financial ratios remained.

Research Results

We deal with big data in the modern world. The features of big data are drastically multiplying, and the increasing dimensionality of data is now a daunting challenge in machine learning and data mining. Moreover, the resultant information might be redundant, irrelevant, and obsolete. The algorithms may lose efficiency as the features increase (Khalid *et al.*, 2014). Feature selection is aimed at improving the classification accuracy and reducing the classification error. Therefore, as the redundant features are reduced, a few features with appropriate information remain and improve the learning process (*e.g.*, further learning precision for classification), decrease computational cost, and enhance the model interpretability (Vieira *et al.*, 2010).

The *k*-nearest neighbors 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 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 used as training data in MATLAB to determine 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 used as the test data in MATLAB to evaluate the algorithms and analyze fraud risk prediction.

Results of Evaluating Proposed Method in Fraud Prediction without Feature Reduction

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

$$Accuracy = \frac{\mathrm{TP + TN}}{TP+TN+FP+FN}$$

(2)

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

(3)

$$Recall \ = \frac{\mathrm{TP}}{TP+FN}$$

(4)

$$F - measure = \frac{2*Precision\ *Recall}{Precision+Recall}$$

(5)

Table1 present brief reports of results obtained from classifiers with all financial ratios (*i.e.*, 96 financial ratios) through the above evaluation criteria after 30 executions with test data.

Table 1. Brief results of Performance Evaluating of Proposed Methods without Feature Reduction

Criterion	Accuracy	Precision	Recall	F-measure
K-NN	0.6620	0.5272	0.6218	0.5706
Bayesian Network	0.6551	0.5198	0.5897	0.5526
SVM	0.6944	0.5690	0.6346	0.6000
Bagging	0.7245	0.6121	0.6474	0.6293

The values extracted for evaluating the results of the proposed classifiers with all financial ratios are low and inappropriate, a fact which indicates that proper features should be selected and used for classification to improve the results.

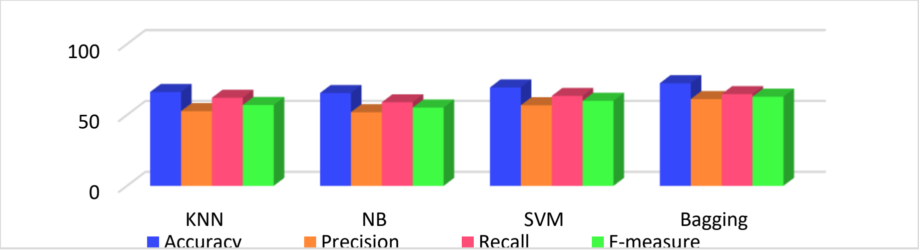

Figure 4 Brief results of Performance Evaluating Proposed Method without Feature Reduction

Results of Analysing Proposed Method without Feature Reduction on Confusion Matrix

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 consists of the following elements:

True Positive (TP): The suspiciously fraudulent financial statements identified correctly

False Positive (FP): The suspiciously fraudulent financial statements identified wrongly as non-fraudulent

True Negative (TN): The non-fraudulent financial statements identified correctly

False Negative (FN): The non-fraudulent financial statements identified wrongly as suspiciously fraudulent

Table 2 present the confusion matrix results obtained from the Proposed Methods without Feature Reduction through test data:

Table 2. Brief results of confusion matrix of test data of the Proposed Methods without Feature Reduction

Algorithm	K-NN	Bayesian Network	SVM	Bagging
TP	97	92	99	101
TN	189	191	201	212
TP + TN	286	283	300	313
FP	87	85	75	64
FN	59	64	57	55
FP + FN	146	149	132	119

Selecting Financial Ratios through PSO

In the second step, the PSO algorithm (*i.e.*, a metaheuristic method) was used in MATLAB to select the best financial ratios from 96 ratios. In this algorithm, the initial population of particles is defined as a collection of *n*-dimensional vectors, the lengths of which equal the number of highly correlated features extracted from the previous step. The positions of particles in the initial population are determined randomly, and the velocity of each particle is initiated with zero by default. The initial particles are evaluated in the first step, and the optimal particles are selected. The rest of the particles are then updated. The positions of the non-optimal particles change with respect to those of the optimal particles, and the convergence speeds of particles are calculated through the fitness function in the next step.

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

In the end, after the algorithm is iterated, the final optimal particle is selected as the near-optimal solution. Table 3 presents the selected features based on the optimal particles in this algorithm for 11 financial ratios.

Table 3. The financial ratios selected by the PSO algorithm

Financial Ratio	Number of Features in 30 Iterations	Financial Ratio	Number of Features in 30 Iterations
Total debts to total assets	26	Net profit to gross profit	17
Working capital to total assets	22	Current assets to current debts	15
Inventory to current assets	23	Cash inventory to current debts	16
Accounts receivable to sales	27	Accumulated profit and loss to equity	23
Accounts receivable to total assets	25	Long-term debts to equity	18
Gross profit to total assets	29		

These financial ratios were selected as optimal features from the highly correlated features. Since the main criterion for evaluating the particles in the PSO is the detection error of financial statements, the features selected by a particle will be more optimal if the fitness function value is smaller for that particle. Figure 5 demonstrates the convergence of the fitness function values on the optimum by the PSO algorithm.

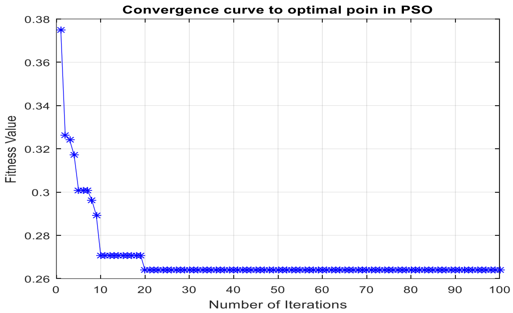

Figure 5. The convergence of the fitness function on the optimum in the PSO algorithm

According to Figure 5, the fitness function values of the PSO algorithm in the feature subset selection problem converged on the optimum with an error rate of zero as the iterations increased. After 100 iterations in this algorithm, the fitness function was obtained 0.2708, and the accuracy of financial ratios selected by the PSO algorithm was 72.92% for training data to detect non-fraudulent financial statements and those suspected of fraud. This algorithm yielded the best financial ratios after 19 iterations at a high speed.

Validity of PSO

The test data were employed to analyze the validity of financial ratios extracted by the PSO algorithm.

Table 4. Validity of the PSO algorithm

Detection Result	Non-Fraudulent	Suspected of Fraud	Total	Precision
Non-fraudulent	240	36	276	84.82%
Suspected of Fraud	27	129	156	

Evaluation Results of Proposed Fraud Prediction Methods through Designated Financial Ratios

Tables 5 to 8 report the brief results of evaluating the proposed classifiers with 11 financial ratios extracted by the PSO algorithm based on four evaluation criteria after 30 iterations with test data.

Table 5. Brief results of evaluating the *k*-nearest Neighbors algorithm with designated financial ratios

Criterion	Accuracy	Precision	Recall	F-measure
Mean	0.7803	0.6804	0.7389	0.7029
Maximum	0.7893	0.6923	0.7500	0.7200
Minimum	0.7638	0.6570	0.7180	0.6765
Standard Deviation	0.0068	0.0100	0.0110	0.0132

Table 6. Brief results of evaluating the Bayesian network with designated financial ratios

Criterion	Accuracy	Precision	Recall	F-measure
Mean	0.7555	0.6484	0.7060	0.6822
Maximum	0.7639	0.6588	0.7179	0.6988
Minimum	0.7407	0.6294	0.6730	0.6532
Standard Deviation	0.0057	0.0073	0.0128	0.0115

Table 7. Brief results of evaluating the support vector machine with designated financial ratios

Criterion	Accuracy	Precision	Recall	F-measure
Mean	0.8060	0.7120	0.7773	0.7432
Maximum	0.8148	0.7236	0.7884	0.7546
Minimum	0.7824	0.6782	0.7500	0.7151
Standard Deviation	0.0076	0.0110	0.0130	0.0098

Table 8. Brief results of evaluating the bagging method with designated financial ratios

Criterion	Accuracy	Precision	Recall	F-measure
Mean	0.8428	0.7685	0.8083	0.7878
Maximum	0.8541	0.7818	0.8269	0.8037
Minimum	0.8217	0.7365	0.7692	0.7616
Standard Deviation	0.0087	0.0122	0.0182	0.0123

Results of Analyzing Models with Confusion Matrix

Table9 report the results obtained from the confusion matrix with the test data regarding the proposed methods and the extracted financial ratios.

Table 9. Brief results of confusion matrix of test data of particle swarm optimization algorithm with Proposed Methods

Algorithm	K-NN	Bayesian Network	SVM	Bagging
TP	117	112	123	129
TN	224	218	229	240
TP + TN	341	330	352	369
FP	52	58	47	36
FN	39	44	33	27
FP + FN	91	102	80	63

Results of Analyzing Precision and Error of Financial Ratios with the Proposed Methods

Table 10 reports the results of analyzing precision and error in prediction of the financial ratios extracted by the proposed classifiers with the test data.

Table 10. Results of analyzing precision and error in prediction of financial ratios extracted by proposed Methods

	Prediction Precision				Prediction Error			
	K-NN	Bayesian Network	SVM	Bagging	K-NN	Bayesian Network	SVM	Bagging
PSO	0.6804	0.6484	0.71120	0.7685	0.3196	0.3516	0.2880	0.2315

AUC of ROC Proposed Method without Feature Reduction

The ROC curve depicts the efficiency of financial ratios extracted by the PSO algorithm with the proposed classification methods based on the values of accuracy, precision, recall, true positive, and false positive. The areas under the ROC curves of PSO–KNN, PSO–NB, PSO–SVM, and PSO–bagging models were reported 79.80%, 80.10%, 83.03%, and 87 %, respectively. These values indicate the efficiency of each model in predicting non-fraudulent companies and those suspected of fraud.

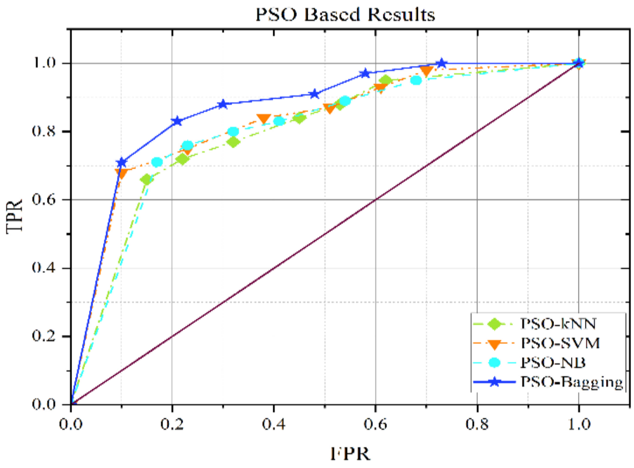

Figure 6. Evaluating the efficiency of financial ratios extracted by the PSO with the proposed Methods

Result Analysis of the First Hypothesis

Table 11 compares the results of evaluating the proposed methods for predicting fraud and classifying companies in terms of four performance evaluation criteria (*i.e.*, accuracy, precision, recall, and F-measure) without feature reduction (96 financial ratios) and after feature reduction (11 financial ratios) with the ratios extracted by the PSO. Since all methods improved significantly after the financial ratios were reduced, the first hypothesis was confirmed. Hence, reducing features (*i.e.*, financial ratios) can be more effective in fraud risk prediction than not reducing features.

Table 12 indicates the consistency between the financial ratios extracted by the PSO algorithm and the financial ratios used in some of the previous studies:

Table 11. Brief results of performance evaluation criteria for proposed algorithms with and without reduction in financial ratios

Criterion	Accuracy	Precision	Recall	F-measure
Performance results of proposed algorithms with all financial ratios				
K-NN	66.20%	52.72%	62.18%	57.06%
Bayesian Network	65.51%	51.98%	58.97%	55.26%
SVM	69.44%	56.90%	63.46%	60%
Bagging	72.45%	61.21%	64.74%	62.93%
Performance results of proposed algorithms with financial ratios extracted by PSO				
K-NN	78.03%	68.04%	73.89%	70.29%
Bayesian Network	75.55%	64.84%	70.60%	68.22%
SVM	80.60%	71.20%	77.73%	74.32%
Bagging	84.28%	76.85%	80.83%	78.78%

Table 12. Results of analyzing the extracted financial ratios in comparison with the previous ratios

Financial Ratio	Previous Studies
Total debts to total assets	Chang *et al.* (2021); Jan (2018); Omar *et al.* (2017); Kamrani & Abedini (2022); Rezaei *et al.* (2020); Kazemi (2016)
Working capital to total assets	Omar *et al.* (2017); Tashdidi *et al.* (2019); Bahmanmiri & Malekian (2016)
Inventory to current asset	Bahmanmiri & Malekian (2016)
Accounts receivable to sales	Omar *et al.* (2017); Kamrani & Abedini (2022); Tashdidi *et al.* (2019)
Accounts receivable to total assets	Chang *et al.* (2012); Kamrani & Abedini (2022); Rezaei *et al.* (2020)
Gross income to total assets	Omar *et al.* (2017); Bahmanmiri & Malekian (2016)
Net income to gross income	Tashdidi *et al.* (2019)
Current asset to current debt	Omidi *et al.* (2019); Jan (2018); Rezaei *et al.* (2020); Tashdidi *et al.* (2019); Kazemi (2016)
Retained earnings and loss to equity	Kazemi (2016)
Long-term debt to equity	Omidi *et al.* (2019); Tashdidi *et al.* (2019)

Result Analysis of the Second Hypothesis

Table 13 reports the results of analyzing the proposed models in terms of four performance evaluation criteria, confusion matrix, precision, prediction error, and ROC. The superiority of the PSO–bagging method (*i.e.,* financial ratios extracted by the PSO and the bagging method) confirmed the second research hypothesis. In fact, using classification algorithms with the bagging method can be more effective in fraud risk prediction than the other classifiers of *k*-nearest Neighbors algorithm, Bayesian network, and support vector machine.

Table 13. Brief results of confusion matrix, precision, prediction error, and ROC

Algorithms	KNN	Bayesian Network	SVM	Bagging
Confusion Matrix (TP + TN)	341	330	352	369
Precision	68.04%	64.84%	71.20%	76.85%
Error	31.96%	35.16%	28.80%	23.15%
ROC	79.80%	80.10%	83.03%	87.00%

Conclusion

Fraudsters use organized but complex schemes to deceive people and organizations; therefore, many costly fraud cases that deceive investors, creditors, and other users and cause serious non-financial harm (*e.g.,* damage to the reputation of accountants) are not detected in a timely manner. This highlights the importance of developing effective methods for financial statement fraud detection. The large number of independent variables that affect the detection of financial statement fraud increases the error rate in detecting fraud. In addition, statistical methods are better at predicting relationships when the data are continuous and linear than in cases where discrete and non-linear data are used (*e.g.,* fraud prediction) (Ranganathan *et al.*, 2017). Accordingly, the authors of this study used the particle swarm optimization (PSO) algorithm to reduce the number of research variables from 96 financial ratios to 11 ratios. Then, several data mining methods, including the k-nearest neighbor (k-NN) algorithm, Bayesian network (BN), SVM, and the bagging method were employed to predict financial statement fraud considering the aforementioned financial ratios. Larose (2005) argues that these methods are among the top ten techniques for discovering unknown relationships and data patterns (Berry & Linoff, 2004). In addition, researchers such as Ali, Lei, Aftabi, Chen, Shogo, Cheng, Gupta, Hedayatollah, Sad Gali, Yao, Kazemi, Kamrani, Rezaie, Tashdidi, Ebrahimi have confirmed the superiority of these methods over other approaches in fraud detection. Financial ratios can help experts accurately predict cases of financial statement fraud. In this study, the PSO algorithm improved the performance of classification algorithms by reducing the number of financial ratios. In addition, the bagging method outperformed other fraud prediction methods with an accuracy of 84.28%, an efficiency of 87%, and an error rate of 23.15%.

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.
- ✓ Legal and monitoring authorities of Iran, researchers, stakeholders, and other users of financial statements can benefit from the proposed PSO–bagging method (*i.e.*, financial ratios extracted by the PSO and the bagging classifier) to predict fraud risk at companies, for it is more efficient in prediction.
- ✓ 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.

References

Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. Expert Systems with Applications, 227, 120144.

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.

Auditing Standards Committee (2015). *Principles and Regulations of Accounting and Auditing: Auditing Standards*, *Audit Organization Publications*, Tehran, Iran

Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

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.

Cheng, C. H., Kao,Y.F., &lin, H. P. (2021). A financial statement fraud model based on synthesized attribute selection and a dataset with missing values and imbalanced classes. Applied Soft Computing 108: 107487.

Chimonaki, C., Papadakis, S., Vergos, K., & Shahgholian, A. (2018, June). Identification of financial statement fraud in Greece by using computational intelligence techniques. In International Workshop on Enterprise Applications, Markets and Services in the Finance Industry .pp. 39-51. Springer, Cham.

Cormen, T. H., Leiserson, C., Rivest, R., & Stein, C. (2001). Advanced Algorithms-CS 6/76101

Corruption Perceptions Index . (2021).https://www.transparency.org/en/cpi

Ebrahimi, M.,& Khajavi, SH. (2017). Modeling Effective Variables in Fraud Detection in Financial Statements through Data Mining Techniques, Financial Accounting Journal, (33): 41–62

El-Shorbagy, M. A., & Hassanien, A. E. (2018). Particle swarm optimization from theory to applications. International Journal of Rough Sets and Data Analysis (IJRSDA), 5(2), 1-24.

Guo,G., Wang, H., Bell, D., Bi, Y., & Greer,K. (2003).KNN Model-Based Approach inClassification, Lecture Notes in Computer Science,Volume 2888.

Gupta, S., & Mehta, S. K. (2021). 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. Global Business Review, 0972150920984857.

Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. Morgan Kaufmann, 340, 94104-3205.

Hidayattullah, S., Surjandari, I., & Laoh, E. (2020). Financial Statement Fraud Detection in Indonesia Listed Companies using Machine Learning based on Meta-Heuristic Optimization. International Workshop on Big Data and Information Security (IWBIS). IEEE. 79-84

Hosseini, S.M., Mahfoozi, Gh., & Kheradyar, S. (2021). Relationship Between Tax Reporting Aggressiveness and Financial Statement Fraud; Accounting and Auditing Research, 13 (50): 163–176.

Ibadin, P. O., & Kemebradikemor, E. (2020). Tax Fraud in Nigeria: A Review of Causal Factors. Journal of Taxation and Economic Development, 19(1), 64-80.

Kamrani,H.,& Abedini, B. (2022). *Developing Fraud Detection Model in Financial Ratios through Neural Network Methods and Support Vector Machine at TSE-Listed Companies*, *Accounting Knowledge and Management Auditing*, (41): 285–314

Kazemi, T.(2016). *Identifying Cases of Fraud Risk in Financial Statements of Iran and Evaluating Fraud Detection Methods*, Doctoral Dissertation, Faculty of Economics and Social Sciences, Shahid Chamran University of Ahvaz.

Khalid, S., T. Khalil, and S. Nasreen.(2014). A survey of feature selection and feature extraction techniques in machine learning. in 2014 science and information conference IEEE.

Kuncheva, L. I. (2014). Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

Larose, D. T. (2005). An introduction to data mining. Traduction et adaptation de Thierry Vallaud.

Lei, Y., Qiaoming, H., & Tong, Z. (2023). Research on Supply Chain Financial Risk Prevention Based on Machine Learning. Computational Intelligence and Neuroscience.

Leung, K. M. (2007). Naive bayesian classifier. Polytechnic University Department of Computer Science/Finance and Risk Engineering, 2007, 123-156.

Nath, S.S., G. Mishra., J. Kar., S. Chakraborty & N. Dey. (2014). A survey of image classification methods and techniques. In 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)IEEE: 554-557.

Occupational Fraud. (2022,2020.2018.2016).A Report To Nations ,https://acfepublic.s3.us-west-mazonaws.com.

Pradhan, A. (2012). Support vector machine-a survey. International Journal of Emerging Technology and Advanced Engineering, 2(8), 82-85.

Ranganathan, P., Pramesh,C.S., & Aggarwal,R. (2017). Common pitfalls in statistical analysis: Logistic regression. Perspectives in clinical research 8(3): 148

Rahnama Roudposhti, F. (2012). *Data Mining and Financial Fraud Detection*, *Accounting Knowledge and Management Auditing*, 1 (3): 17–33

Rastatter, S., Moe, T., Gangopadhyay, A., & Weaver, A. (2019). Abnormal Traffic Pattern Detection in Real-Time Financial Transactions (No. 827). EasyChair.

Rezaie, M., Nazemi Ardakani, M., & Naser Sadrabadi, A. (2021). Predicting Financial Statement Fraud with CRISP Approach; Management Accounting and Auditing Knowledge, 10 (40): 135–150.

Rezaei, M., Nazemi Ardakani, M.,& Naser Sadr Abadi, A. (2020). *Fraud Detection in Financial Statements through Audit Reports of Financial Statements*, *Management Accounting Journal*, (45): 141–153

Sadgali, I., N. Sael & F. Benabbou. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia computer science 148: 45-54.

Shinde, A., Sahu, A., Apley, D., & Runger, G. (2014). Preimages for variation patterns from kernel PCA and bagging. Iie Transactions, 46(5), 429-456.

Tashdidi, E., Sepasi, S., Etemadi, H., & Azar, A. (2019). *Proposing a Novel Approach to Fraud Prediction and Detection in Financial Statements through Bees Algorithm, Accounting Knowledge Journal,* 12 (3): 139–167.

Umar, H., Purba, R. (2020), "HU Model: Incorporation of Fraud Star in Detection of Corruption", International Journal of Economics and Management Studies, 13(6), PP. 234-265.

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(4), 2714-2723.

Wang, J., Cao, Y., Li, B., Kim, H. J., & Lee, S. (2017). Particle swarm optimization based clustering algorithm with mobile sink for WSNs. Future Generation Computer Systems, 76, 452-457.

Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. Decision support systems, 57, 77-93.

Xiuguo, W., & Shengyong, D. (2022). An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. IEEE Access, 10, 22516-22532.

Yao, J., Pan, Y., Yang, S., Chen, Y., & Li, Y. (2019). Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: a multi-analytic approach. Sustainability, 11(6), 1579.

Yao, J., Zhang, J., & Wang, L. (2018, May). A financial statement fraud detection model based on hybrid data mining methods. International Conference on Artificial Intelligence and Big Data (ICAIBD). pp. 57-61. IEEE.

Yingquan W., Ianakiev,K., & Govindaraju,V. (2002). Improved k-nearest neighbor classification ",Pattern Recognition 35.

Youkhneh Alghiani, M., Bahri Sales,J., Jabarzadeh Kangarlouei,S.,& Zavari Rezaei, A. (2021). *Explaining Financial Tax Cross Reporting of Companies: Hybrid Method of Classic Data Mining, ANFIS, and Metaheuristic Algorithms, Empirical Studies of Financial Accounting,* 18 (71), 89–111.