Journal of Optimization in Industrial Engineering
Vol.14, Issue 2, Summer & Autumn 2021, 183-201
DOI: 10.22094/JOIE.2020.1877455.1685

A Hybrid Model Based on Machine Learning and Genetic Algorithm for Detecting Fraud in Financial Statements

Akbar Javadian Kootanaee [a], Abbas Ali Poor Aghajan [b], Mirsaeid Hosseini Shirvani [c]

[a] *Department of Accounting, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran,*
[b] *Department of Accounting, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran,*
[c] *Department of computer engineering, Sari branch, Islamic Azad University, Sari, Iran.*
Received 27 August 2019; Revised 03 August 2020; Accepted 05 August 2020

Abstract

Financial statement fraud has increasingly become a serious problem for business, government, and investors. In fact, this threatens the reliability of capital markets, corporate heads, and even the audit profession. Auditors in particular face their apparent inability to detect large-scale fraud, and there are various ways to identify this problem. In order to identify this problem, the majority of the proposed methods are based on existing algorithms and have only attempted to identify human or simple data mining methods that have high overhead and are also costly. The data mining methods presented so far have had high computational overhead or low accuracy. The purpose of this study is to present a model in which an improved ID3 decision tree with a support vector machine is used as a hybrid approach and also to improve the performance and accuracy, genetic algorithm and multilayer perceptron neural networks are applied. More efficient feature selection has been used to reduce computational overhead. The tree proposed in the proposed method has the lowest depth possible and therefore has high velocity and low computational overhead. For this purpose, the financial statements of 151 listed companies in Tehran Stock Exchange during 2014-2015 were surveyed and 125 financial ratios were extracted using ANOVA test, 23 fraud related ratios were selected as model input data. The proposed model has a high accuracy of about 80% of prediction accuracy compared to similar models.

Keyword: Support vector machine; Improved decision tree; Fraud detection; Classification.

1. Introduction

Financial statement fraud is increasingly a serious problem for businesses, governments and investors. In fact, this threatens the reliability of capital markets, corporate heads, and even the audit profession. Auditors in particular face their apparent inability to detect large-scale fraud. In recent years, the US financial markets have been seriously damaged by numerous disclosures of fraudulent practices by some companies. WorldCom, Enron, Adolfia, Global Cracking and Tico are just a few of the financial statement scandals that have fluctuated the stock market and eroded public confidence (Vakili fard & Ahmadi, 2010).

Financial statement fraud is increasingly a serious problem for businesses, governments and investors. In fact, this threatens the reliability of capital markets, corporate heads, and even the audit profession. Auditors in particular face their apparent inability to detect large-scale fraud. Hundreds of millions of dollars in monetary judgments are typically made against companies with formal public accounting services.

From the audit perspective, fraud is a very serious issue as fraud is often associated with trying to conceal, distort and mislead users of the audited records and reports. Attempts to misrepresent information can also occur at the management level, as this is widely accepted following the collapse of large corporations.

Data mining is not just limited to social interactions, science and engineering, but it is also used in recommender systems, financial systems, anti-spyware and more. Although data mining methods have many applications in different sciences. So far the rate of adoption of these methods by academics and control organizations to detect fraud has not been significant. In the early stages of this research, problem expression and introductory concepts in the domain are examined. Based on recent events and observations, information theft and financial scandals, it can be argued that intra-organizational fraud is likely to occur in any company or entity, whether commercial or non-commercial, and specific to a particular level or category of that set. (Andon, Paul et al., 2015), (Lookman & Selmin Nurcan,2015).

*Corresponding author Email address: abbas_acc46@yahoo.com

2. Literature Review

2.1. Fraud

According to the Auditing Standards Committee (2015), "fraud" is any deliberate or fraudulent act by one or more managers, employees, or third parties to gain an unjust or illegal advantage. Although fraud has a broad legal meaning, what concerns the auditor are fraudulent practices that lead to significant misstatement of the financial statements. The purpose of some frauds may not be to distort financial statements. Auditors do not make legal judgments about fraud. The committee argues that fraud involving the involvement of one or more of the unit managers is referred to as "fraud of managers" and fraud that is only perpetrated by the staff of the unit being referred to as "employee fraud". In both cases, there may also be collusion with third parties outside the unit concerned (Iranian Audit Organization, 2015).

Rezai and Riley (2010) in their book divided fraud into two categories of management fraud and employee fraud, and below provide a further classification of these two types of fraud as illustrated in Fig 1. Fraud can be divided into several types, the most common being confiscation of assets and financial errors. Confiscation of assets often involves employee fraud, including embezzlement, cash or inventory robbery, and payroll fraud; financial errors are regarded as fraudulent financial statements, which are often the responsibility of management. The US Department of Justice defines corporate fraud in three broad areas: accounting or financial fraud, fraud, and deviant behavior. Accounting fraud involves distorting financial information through accounting or misleading investors. The most common accounting schemes include asset sales, side trading, exchange trades, investment costs, quick cash earnings and deferred expenses (Rezaee & Riley, 2010).

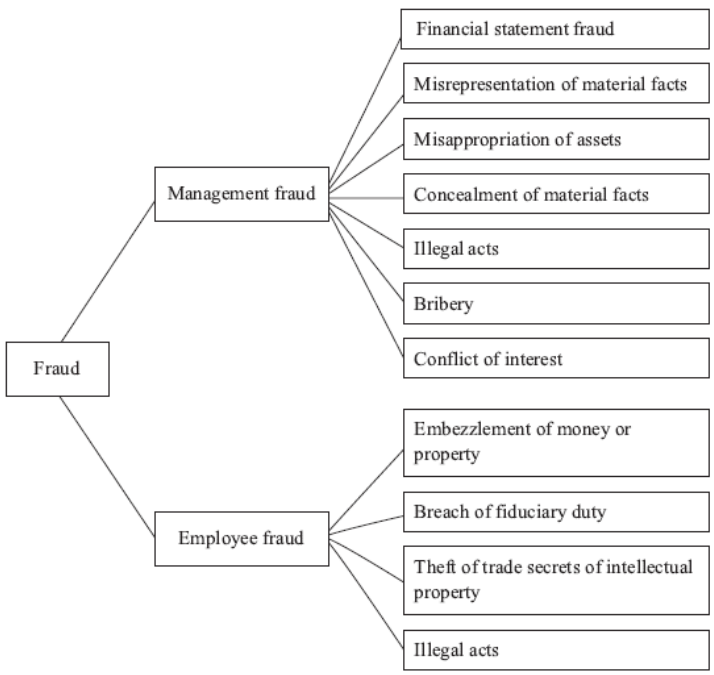

Fig 1. Fraud Types (Rezaee & Riley, 2010)

2.2. Machine Learning

Machine learning is how to write a program that will improve your performance through the learning experience. Learning can change the structure of the program or data. Machine learning is a relatively new field in computer science that is currently undergoing growth. Machine learning is a very active field of research in computer science (Ziming Yin et. al. 2014). There are various sciences related to machine learning including artificial intelligence, psychology, philosophy, information theory, statistics and probability, control theory, and so on.

A computer program from experience E has learned about the task of T if its performance improves if measured by the P criterion after that experience.

The reasons for using machine learning to solve problems include:

- Large amounts of data may contain important information that humans cannot detect (data mining).
- While designing a system, all of its features may not be known while the machine can learn them while working.
- The environment may change over time. The machine can adapt to them by learning about these changes.

Some of the machine learning applications include robots control, data mining, speech recognition, text recognition, internet data processing, bioinformatics, computer games, and thousands of other examples. The basics of evaluating machine learning algorithms are classification accuracy, solution accuracy and quality, and performance speed. Machine learning is divided into two general categories of supervised learning and supervised learning (Senthil Kumar et. al. 2013)

2.3. *Genetic algorithm*

The scope of the genetic algorithm is vast and the use of this method in optimizing and solving problems is expanding with the advances of science and technology. Genetic algorithm is one of the evolved computing subsets that is directly related to artificial intelligence. In fact, genetic algorithm is one of the subsets of artificial intelligence. The genetic algorithm can be called a general search method that mimics the laws of natural biological evolution. In each generation, better approximations of the final answer are obtained through the selection process proportional to the value of the responses and the reproduction of the responses selected by operators mimicking natural genetics. This process makes the new generations more adaptable to the problem conditions. (http://hkamal.persiangig.com/document/genetic, 2018). Note that, genetic algorithm can be utilized in both single objective (Hosseini Shirvani & Babazadeh Gorji, 2020; Hosseini Shirvani, 2018a; Hosseinzadeh & Hosseini Shirvani, 2015) and multi-objective optimization problems (Farzai et al., 2020; Hosseini Shirvani et al., 2018; Hosseini Shirvani, 2018b). In this paper, the signle objective application of genetic algorithm is engaged.

2.4. *Multilayer perceptron network*

In this type of network the connection is only from i to i + 1 and there is no reverse. The above network is actually created by the interconnection of three single layer perceptron networks. One is called the output layer and the other two are called the middle layers. The first layer outputs form the second layer input vector, and so the second layer output vector constructs the third layer inputs, and the third layer outputs form the true Rattan network response. In other words, the signal flow process in the network takes place in a predetermined direction (from left to right from layer to layer). Each layer can have a number of different neurons with different conversion functions, meaning that the models of neurons in the layers can be considered differently. During the training of the MLP network using BP learning algorithm, the calculations are first performed from the network input to the network output and then the calculated error values are propagated to the previous layers (https://chistio.ir ،2019). Functional signals and signals that go back (right to left) are the error signals that move along the path (from left to right of the network). Fig 2 illustrates these two releases.

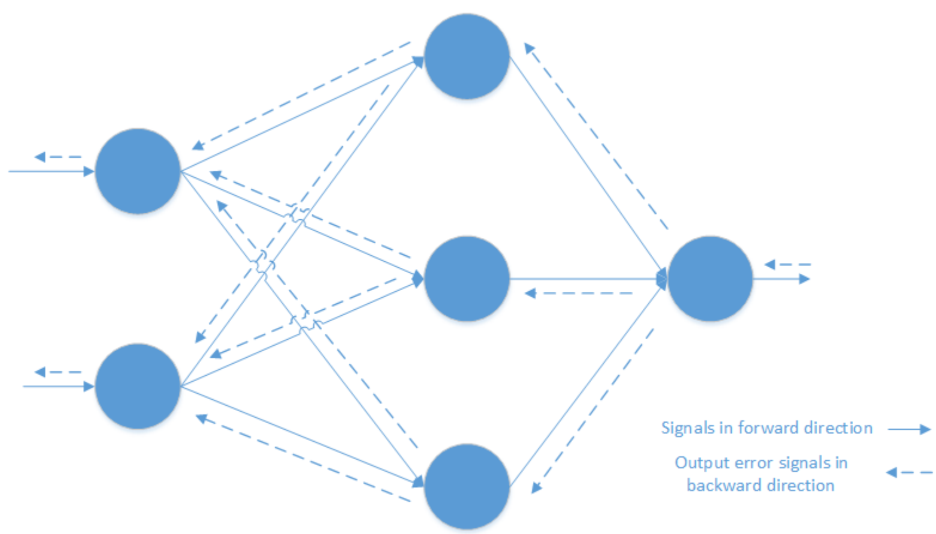

Fig 2. Signal propagation in bp algorithm

2.5. *Decision tree*

Decision tree structure in machine learning is a predictive model that contributes the observed facts about a phenomenon to inferences about the purpose value of the phenomenon. The machine learning technique for deriving a decision tree from data is called decision tree learning, which is one of the most common data mining methods (Kantesh Kumar, et. al., 2014).

Each node corresponding to a variable and each arc to a child represent a possible value for that variable. A leaf node, with the values of the variables represented by the path from the tree root to that leaf node, represents the predicted value of the target variable. A tree represents a structural decision tree whose leaves represent clusters and branches represent seasonal combinations of traits that result in these clusters (Ojeme Blessing et. al, 2014). Learning a tree can be done by subdividing a resource set into subsets based on a trait value test. This process is repeated recursively in each subdirectory resulting from the separation. The return operation is complete when the separation is no longer beneficial or one class can be applied to all the samples in the subgroup.

Decision trees are capable of generating human-readable descriptions of relationships within a dataset and can be used for classification and prediction tasks. This technique has been widely used in a variety of fields such as plant disease diagnosis and customer marketing strategies. (Ojeme Blessing et. al, 2014). This decision structure can also be introduced in the form of mathematical and computational techniques that help describe, classify and generalize a set of data.

Types of Tree Properties Decide on two types of batch and real traits, which are batch traits that accept two or more discrete values (or symbolic traits) while the true traits derive their values from the set of real numbers.

2.5.1. Development of decision trees with decision graphs

Decision graphs are generalizations of decision trees that have decision leaves and nodes. One feature that distinguishes decision graphs from decision trees is that decision graphs can be linked. Transplantation is a condition in which two nodes have a common child, and this condition represents two subunits that have common characteristics, and therefore are considered a set (Dionysios, 2018). In the decision tree, all paths go from the root node to the leaf node with the AND compound. In a decision graph it is possible to use seasonal combinations or ORs to link two or more paths together.

The way objects are categorized in decision graphs is the same as the one used in decision trees. Each decision tree and decision graph define a category (a partition of object space into separate categories). The set of functions represented by a graph is exactly the same as the set represented by a tree. However, the set of categories that are included in the definition of a decision function is different. For example, the classification for function $(A \wedge B) \vee (C \wedge D)$ is different. The graph and the corresponding decision tree of this function are shown in Fig. (3). The decision tree divides the object's space into seven categories, while the graph divides the decision space into two categories (Dreżewski, et. al. 2015).

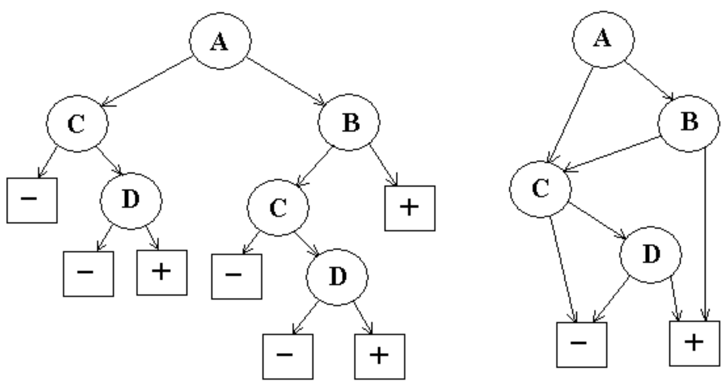

Fig 3. An example of a decision graph (Dreżewski, et. al. 2015).

2.6. Research background

Sudan Chen (2016), in a study entitled "Detecting Fraudulent Financial Statements Using a Combined Data Mining Approach", with the aim of establishing a valid fraud detection model in the financial statements of companies listed on the Taiwan Stock Exchange between 2002 and 2013, both fraudulent financial statements and tested non-fake. In the first step, two decision tree algorithms, including (CART) and (CHAID), are used to select the main variables. The second step involves combining CART, CHAID, Bayesian network, support vector machine, and artificial neural network to develop fraud detection models. Based on the results, the detection performance of the CHAID-CART model with the overall accuracy of 87.97% is the most effective model.

Kim et al. (2016), in their research, “Detecting fraudulent financial misstatement using cost-sensitive multilevel training” using polynomial logistic regression, support vector machine and Bayesian networks, as tools Predictions to detect and classify misrepresentations based on fraudulent intentions extended the classifier to three levels. They evaluated aspects of previous research to detect fraudulent intentions and the implications of false statements. Aspects such as short-term profit ratio and firm performance scale indicate potential

discriminatory potential.
Larry Dashtbeaz et al. (2015), in a study entitled “The Search and Discovery Process for Financial Statement Fraud,” gave an overview of the data mining processes used to detect financial fraud, particularly corporate financial statement fraud. In their research results, they stated that the most important methods used to detect financial fraud include logistic regression, neural networks, Bayesian inference networks, and decision trees, which are important solutions to the inherent problems of identifying and classifying data.
Mohammad Youssef et al. (2015), in a study entitled “Using Fraud Models in Malaysian Stock Exchange Companies” investigated the possibility of fraudulent financial statements in Malaysian Stock Exchange companies using fraud triangle model, rhombus fraud model and pentagonal fraud model. In this study, the fraud risk factors derived from these fraud models provided a new perspective on the discovery of fraud in the financial statements of Malaysian companies. The results also introduce new measurable criteria and new fraud risk factors such as greed and ignorance.
In their research, "Discovery of Financial Statements Using Data Mining Technique and Performance Analysis", Tangot and Kolkarni (2015) investigated and concluded the application of two methods of data mining called "Key Classification" and "MLFF" algorithm. That the information contained in the financial statements contains fraudulent indicators. In addition, a relatively small number of financial ratios largely determine the classification results. They also concluded that the neural network had a higher accuracy than both other models.
In a study entitled "The Role of Auditors in the Prevention, Detection and Reporting of Fraud in the UAE", Holboni (2015) identifies the processes that internal and external auditors follow to detect fraud during an audit. The sample is 53 auditors in the UAE. The results show that the responsibility for identifying fraudulent incidents with internal auditors is somewhat stricter than that of internal auditors. Overall, the results of this study indicate that it is the responsibility of detecting fraud and reporting it to internal auditors and that external auditors should also increase their search for fraud detection and disclosure.

3. Metodology

There is no specific theoretical framework for identifying and classifying entities into fraudulent or healthy companies in financial reporting. According to the Iranian auditing standards Nos. 240 and 450, the criteria and criteria for making a misstatement can be set on the basis of the amount and content of the characteristics of the financial statements data audited by the auditors or the auditing organization. Therefore, according to previous researches such as Daghmeh Qi Firouzjaie's (1393), Etemadi and Zalaghi (1388), Haghighi and Boroujeri (1388), and Maham et al (1391) class directions to classify sample companies as fraudulent and healthy, the following criteria must be met for at least three years (2014-2015) in the financial statements of the fraudulent companies and the following three conditions are classified as fraudulent companies. These are:

1. Unacceptable audit opinion
2. Tax differences with tax area according to Income tax statement and tax file and paragraph of audit report
3. Existence of significant adjustments and restated financial statements.

The reasons for choosing these criteria are that in the first criterion - the existence of significant fraud, it can give rise to unacceptable commentary and the second criterion in tax disputes is largely due to the misinterpretation of tax laws and the incorrect application of the relevant clauses in certain tax laws. And maintaining liquidity and other potential violations. For the third criterion, misstatement and manipulation of items, especially profit and loss items in the preceding years, may give rise to the re-submission of financial statements and a reason for the likelihood of fraud in the financial statements.
Thus, by the above criteria, first a list of listed companies in Tehran Stock Exchange that committed financial statements fraud between 2014 and 2016 is prepared and the number of fraudulent companies is determined based on the availability of corporate information. Be. They then determine the number of healthy companies in the same time range and then select random samples based on simple random sampling.
It is not possible to match the companies of the two groups in terms of industry because there is no industry or similar industry that has enough of both fraudulent and healthy companies. Thus, sampling will use all the companies listed on the Tehran Stock Exchange, although the goodness of the industries is that the generalizability of the model increases. Only companies are matched for the financial year.
The independent variables in this study are financial ratios, so that by studying and researching the required financial ratios, 125 financial ratios were selected. But in order to avoid high correlations between some ratios and failure to provide similar information, ratios with high correlations were identified and eliminated by T-test. This combination of correlation analysis and T-test resulted in the final selection of 54 independent variables that provide meaningful and non-overlapping information.
Then, these variables are extracted from the financial statements of fraudulent and healthy firms in the studied years and are selected by linear regression model of variables that have significant correlation with fraudulent financial statements and used as input variables of machine learning models and genetic algorithm.
Table 1 reports the mean, standard deviation and ANOVA

test for proportions of fraudulent and non-fraudulent companies. Univariate tests refer to several variables that may be useful in detecting fraudulent companies. Of the 54 variables tested, 23 variables that are significant at the 1 to 5% level are summarized in the table 1 and the other variables were not significant.

Table 1
Mean, standard deviation and ANOVA test for proportions of fraudulent and non-fraudulent companies

No.	Variables	fraudulent		non-fraudulent		ANOVA test	
		Mean	Standard deviation	Mean	Standard deviation	F. test	Prob
X3	Liabilities/total assets	0.85833	0.44489	0.65925	0.3841	2.56392	0.0424
X5	Current ratio	1.21892	0.61016	1.38891	1.63254	28.4676	0.000
X6	Quick ratio	0.60161	0.98004	0.33792	0.46287	17.5058	0.000
X17	Log (CGS)	6.04435	0.73651	5.97154	0.57365	3.72523	0.0114
X18	Net profit/total assets	0.0649	0.15732	-0.04232	0.21988	3.562	0.0005
X20	Net profit/ cost of goods sold	0.01673	0.38055	-0.27322	0.79488	4.73992	0.0018
X21	cost of goods sold/total assets	0.9863	0.77065	0.74075	0.47496	4.04693	0.0458
X24	Operating Profit/sale	-0.09499	0.5091	0.08085	0.2547	9.08334	0.003
X25	Earnings before interest and taxes/sale	-0.2684	0.79852	0.03012	0.38876	10.8867	0.0012
X27	Gross profit/total assets	0.07984	0.13457	0.15411	0.13015	10.8077	0.0012
X28	Earnings before interest and taxes/total assets	-0.03922	0.22325	0.07747	0.17022	13.4629	0.0003
X32	Earnings before interest and taxes/current liabilities	0.08746	0.48497	0.25052	0.46879	4.01394	0.0466
X33	(Current assets - Inventory)/ current liabilities	0.71192	1.16493	0.39511	0.47679	6.68665	0.0105
X34	Inventory/ current liabilities	0.67699	0.61788	0.82381	0.55213	4.0639	0.0244
X35	Cash/ total liabilities	0.08976	0.20681	0.08966	0.20554	3.71495	0.0188
X38	Current liabilities / total assets	0.73133	0.42745	0.57907	0.32242	6.33887	0.0127
X39	Capital / total assets	0.14167	0.44489	0.34075	0.3841	8.36056	0.0043
X42	Inventory/ sale	0.85829	0.91738	0.57365	0.52389	6.49386	0.0117
X43	Accounts Receivable/ sale	0.68403	0.64249	0.43578	0.63394	5.10144	0.0251
X44	Sale / fixed assets	3.94599	3.66542	5.44689	6.05732	5.78711	0.0021
X47	cost of goods sold/sale	0.89015	0.18465	0.82466	0.1531	5.49606	0.0202
X48	Operating costs/sale	0.08085	0.34518	0.01235	0.04305	4.98701	0.0268
X53	Inventory/ current assets	0.54712	0.27773	0.69299	0.39575	5.24274	0.0232

Significant differences in mean values between fraudulent and non-fraudulent companies and high statistical significance ($p < 0.000$) indicate that these ratios are related to fraudulent financial statements. The proposed model has several steps which are as follows:

1. data pre-processing;
2. data transfer;
3. Selecting effective features using genetic algorithms and neural networks;
4. Training and calculation of weights of decision tree algorithms and support vector machine;
5. Making a Decision Tree;
6. Transform the decision tree and optimize it;
7. Making a support vector machine.

The general flowchart of the research method can be seen in Fig. 4.

3.1. Pre-processing data

Initially the dataset is collected and prepares and preprocesses the data. Different methods are used in data preparation and preprocessing. First, some properties have unique values. These attributes cannot create useful knowledge in the dataset. Therefore, this feature set must be deleted from the data. For example, we can mention the name and surname. Some transactions may also have large amounts of missing data. Therefore, these transactions should also be removed from the dataset. On the other hand, some attribute values may have noise and missing values, so these values should also be corrected in the dataset. The next step is to use the anomaly detection tool. Data outside the data set is identified and deleted. In order to work on the data as input, properties must be extracted from them. Typically, some pre-processing operations are performed on the data before selecting and extracting features.

3.2. Data transfer

In this section, the data is in the right domain. That is, the data must be transferred to the suffering specified in the system, and the data outside the suffering is problematic data and must be deleted. The data should be in the right range, meaning that for example if there is an age field, someone between the ages of 55 and 70 should be very old in the system, which will be automatically completed from the data set.

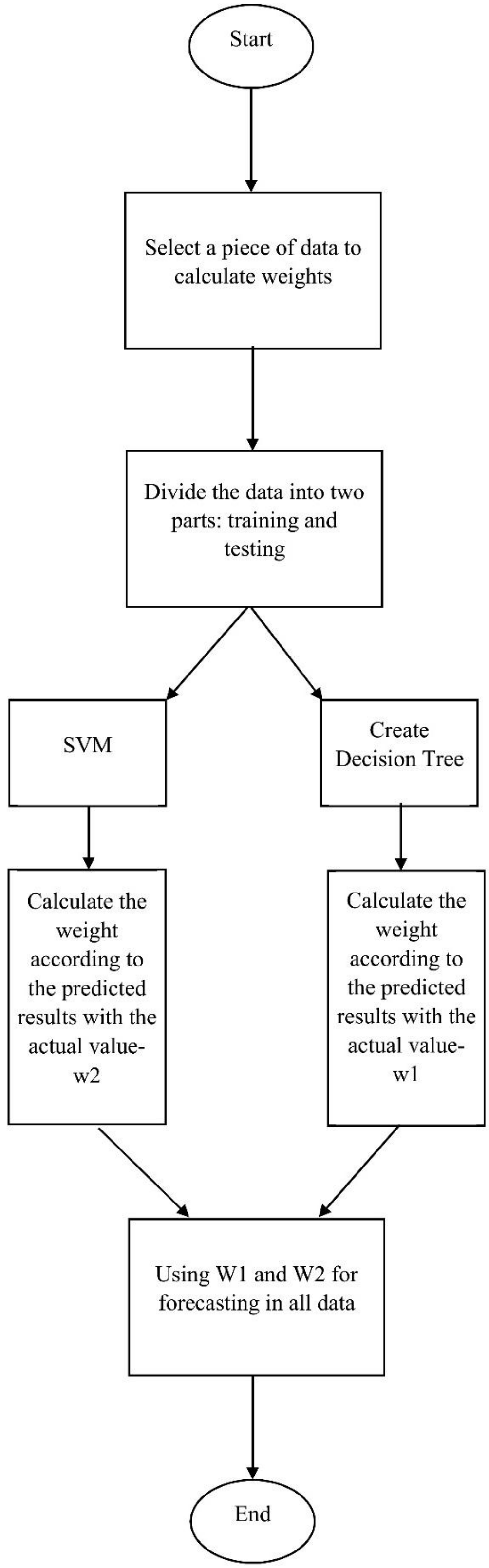

Fig 4. Proposed research model

3.3. Selecting effective features using genetic algorithms and neural networks

Genetic algorithms, with considerable capability to derive meaning from complex data, can be used to extract patterns and identify methods that are very complex and difficult for humans and other computer techniques to be aware of. Genetic algorithm can be used for classification as one of the data mining tools. Here, we try to use the perceptron neural network algorithm, optimized by genetic algorithm, to identify more effective features, and thus identify more effective financial ratios for corporate fraud. In fact, this part of the strategy outlined in the paper presented by Ledsa et al. (2008) is used here.
Determining the network structure is one of the influential steps in how to train the network, although the number of high neurons in the network increases its complexity and the network may be over-fitted which will affect the network's predictability. Therefore, for each training algorithm the number of neurons increased from one to four. The data were thermalized in the range of 1 to 2 and the tansig transition function was used. The two most commonly used optimization algorithms based on biological transformations are Genetic Algorithm, Population in Genetic Algorithm including Possible Responses. In the form of an array of chromosomes. Grid weights were optimized by genetic algorithm so each population was randomly assigned as grid weight. The MSE function is introduced as a cost function, and the population chromosomes are then arranged to achieve the lowest cost function. A certain number of better members are transferred to the next generation at the lowest cost. At this stage, three genetic algorithm operators (selection, intersection and transformation) are activated to generate the next generation population. This cycle continued until the desired solution was achieved to obtain the desired network weights.
Finally, in this section, 23 financial ratios are examined and, if possible, their number is reduced, thereby reducing computational overhead as well as noise, if possible. This process is illustrated in Fig 6. For ease of operation, each ratio using Att and a specified number is shown in Fig. 5.

Att1	:	Liabilities/total assets
Att2	:	Current ratio
Att3	:	Quick ratio
Att4	:	Log (CGS)
Att5	:	Net profit/total assets
Att6	:	Net profit/ cost of goods sold
Att7	:	cost of goods sold/total assets
Att8	:	Operating Profit/sale
Att9	:	Earnings before interest and taxes/sale
Att10	:	Gross profit/total assets
Att11	:	Earnings before interest and taxes/total assets
Att12	:	Earnings before interest and taxes/current liabilities
Att13	:	(Current assets - Inventory)/ current liabilities
Att14	:	Inventory/ current liabilities
Att15	:	Cash/ total liabilities
Att16	:	Current liabilities / total assets
Att17	:	Capital / total assets
Att18	:	Inventory/ sale
Att19	:	Accounts Receivable/ sale
Att20	:	Sale / fixed assets
Att21	:	cost of goods sold/sale
Att22	:	Operating costs/sale
Att23	:	Inventory/ current assets

Fig. 5. List and title of the ratios used

In the proposed method, first, the feature selection task is performed and the more effective features are selected in the proposed method, which for the input dataset can be seen in Fig. 6, which is the effect of the features that are sorted.

Att1 : 0
Att16 : 0
Att15 : 0
Att14 : 0
Att13 : 0
Att23 : 0
Att19 : 0
Att20 : 0
Att18 : 0
Att21 : 0
Att7 : 0
Att22 : 0
Att4 : 0
Att3 : 0
Att2 : 0
Att17 : 0
Att9 : 0.0905061414027294
Att6 : 0.0949449035096511
Att12 : 0.100474368655747
Att11 : 0.104421869607332
Att5 : 0.105127953476521
Att8 : 0.114018997382377
Att10 : 0.120272737712267

Fig 6. The order of influence of the characteristics of the financial ratios used

Considering the above table, we can conclude that the financial ratios of zero coefficient have no effect on output, while the financial ratios such as pre-interest income and sales tax, net interest income, pre-interest income and tax liability. Current earnings before interest and taxes on total assets, net profit on total assets, operating profit on sale and gross profit on total assets have the greatest impact on whether or not the company is fraudulent. Here we can eliminate ineffective financial ratios so that the algorithm is able to perform data mining more quickly, and in this section we can see which ratios have the greatest impact on whether or not they are fraudulent, and what drives them. It's about giving the company a cheat. That is, the influential parameters can be identified here and more attention is paid to identifying the fraudulent companies and making important decisions.

3.4. Training and calculation of weights of decision tree algorithms and support vector machines

At the beginning of the work, a percentage of the data set used to calculate training and weight calculation is used. In this section, we intend to use a hybrid algorithm that uses two decision tree algorithms and a support vector machine. The two algorithms each have a share of the final answer that increases the accuracy of the system. An illustration of this step can be seen in Fig. 7.

As can be seen in Fig. 7, the proposed algorithm first calculates weight using a dataset, so that each algorithm is trained using 70% of the existing dataset and trained with The remaining 30% is tested using the score given by the number of correct answers, or weighted, until the weight of the next step is used to calculate the weight for the output of each algorithm. Architecture can also be seen after calculating the weight divided by the number of correct answers to the total number of barley The water is conjectured, and the effect of each algorithm on the final output can be better understood. In this method, after calculating weights, for each record, a prediction is made by the support vector machine and by the optimal decision tree that the predicted value must be multiplied by the weight of the algorithm and the final output of the prediction algorithm is equal. By summing the results of each algorithm, multiply the weight of that algorithm so that the final result is obtained and the classification is correct.

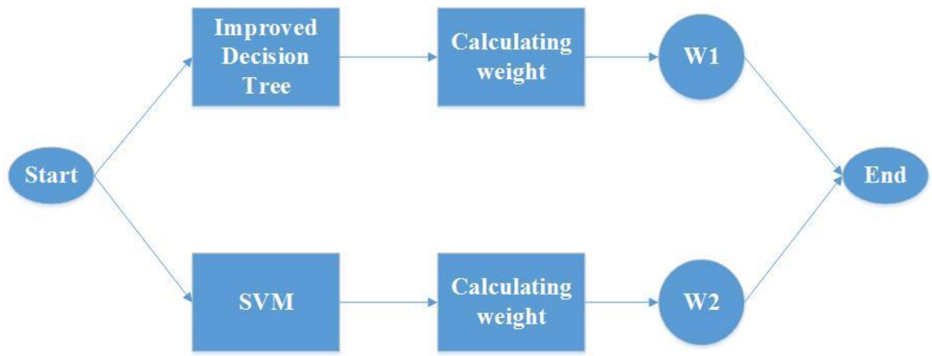

Fig 7. Architecture used in weight calculation

3.5. *Making a decision tree*

This part of the decision tree is used. The decision tree is a tree from which each branch is chosen. That is, one can choose from the branches that are connected to that node to move from the root node to the lower node. At the end, each end node or so-called node leaves a decision leaf. Each branch up to the leaf has a scenario that makes a decision. In this study, the proposed model based on the improved ID3 decision tree is used which results in its high speed of operation. The ID3 tree is a decision tree that also has learning and was first introduced by Ross Quinlan. The idea of the ID3 algorithm is to build a top-down decision tree in which the node is selected by a greedy search through a set of attributes. Here we used a special template to be able to find the most useful attribute among the traits that is most useful in classifying. To make a useful classification for the learning set, the number of questions has to be reduced or the depth of the decision tree can be said to be reduced. Therefore, this part requires the function to be able to perform the most balanced division, in which case the tree depth is greatly reduced and the nodes are balanced in the tree.

Consider a table that contains attributes and a class of attributes. This homogeneous table is called if it contains only one class. If a table has multiple classes, then it is called heterogeneous. There are many functions such as entropy, gini index and classification error to measure homogeneity. Entropy is used here.

$$\text{Entropy} = \sum_{j} -p_j \log_2 p_j \qquad (1)$$

The entropy of a table is zero because its probability is equal to one (tents have one class). The entropy reaches its maximum when all classes in the table have equal probability. Entropy can be considered as a criterion for measuring irregularity. The more regular the set is, the less varied it is, the less entropy it is, and the less irregular it is, and vice versa. Of course here, since we did an elementary classification in the previous step, the chaos is also low, which in turn speeds up our proposed approach because it will reduce the depth of the decision tree and whatever The deeper the tree, the faster the decision is made.

In this section we used entropy to obtain the irregularity for each attribute. We used formula (1) to select an attribute in the decision tree that is higher in rank than the other attributes in some way more important than the other attributes. According to this formula we calculate the entropy of all the attributes in the set S and subtract the value of the attribute A in the set. Set A is the set of all attributes selected by the father so far in a particular path.

$$G(S, A) = Entropy\,(S) - \sum_{v \in Values(A)} Entropy\,(v) \qquad (2)$$

For better understanding of formula (2), we can see Fig 6. As can be seen in Fig 8, we need to entropy all of the attributes from the entropy of the selected attributes to reduce the path here (that is, G (S, F) = E (S) - (E (A) + E (B) + E (E) + E (F)). It should be noted, however, that we need to calculate the attributes A used so far in the set A plus the attribute we want to assign. After this, from this set of residuals we calculated the formula for each (2), choosing the attribute with the highest G. In this case, if two attributes had equal G and the probability of this occurrence is not low, then we should add two or any number of attributes that have the highest value G and are equal to the corresponding node, if, for example, in a node The two attributes had equal G, then we added the two nodes to the corresponding node of each of the two children and each of these attributes were considered as a child of this node and then followed the algorithmic procedure for each of these nodes. We continue. For example, in Fig. 6 we can see that the value of G is equal to the attributes B, C and D, so all these attributes are on the same level.

This allows us to find the attributes that have the most entropy because these attributes have the greatest impact on our final decision. This process of moving forward in the decision tree continues until there is no trait left in any other path. In this case, the decision tree is completely built and finished.

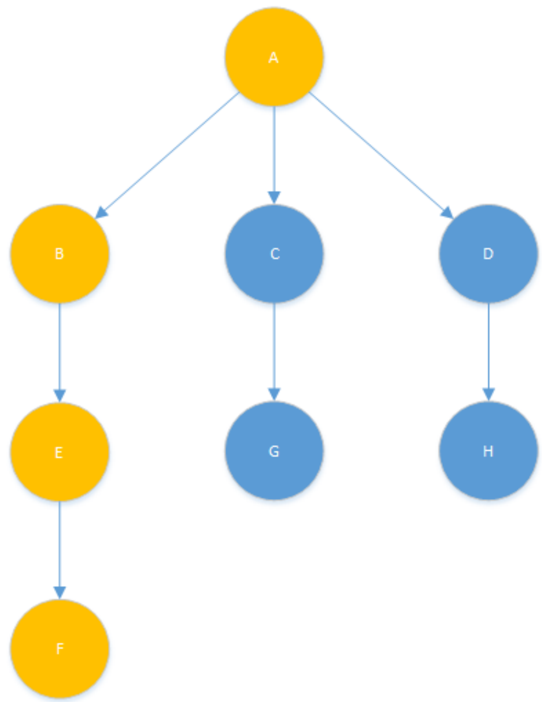

Fig. 8. An example of selecting attributes

3.6. Improving the decision tree

In order to extract conditions from this tree as if… then…. Where associative rules can be helped, we have transformed the tree into a graph that in many cases becomes a tree, and even if it looks like a graph, it will still be for us. The tree is considered so something can be said between the graph and the tree.
In this section we merge the nodes at each level and add their children to this merge node. This can be seen in Fig 9 to better understand it.

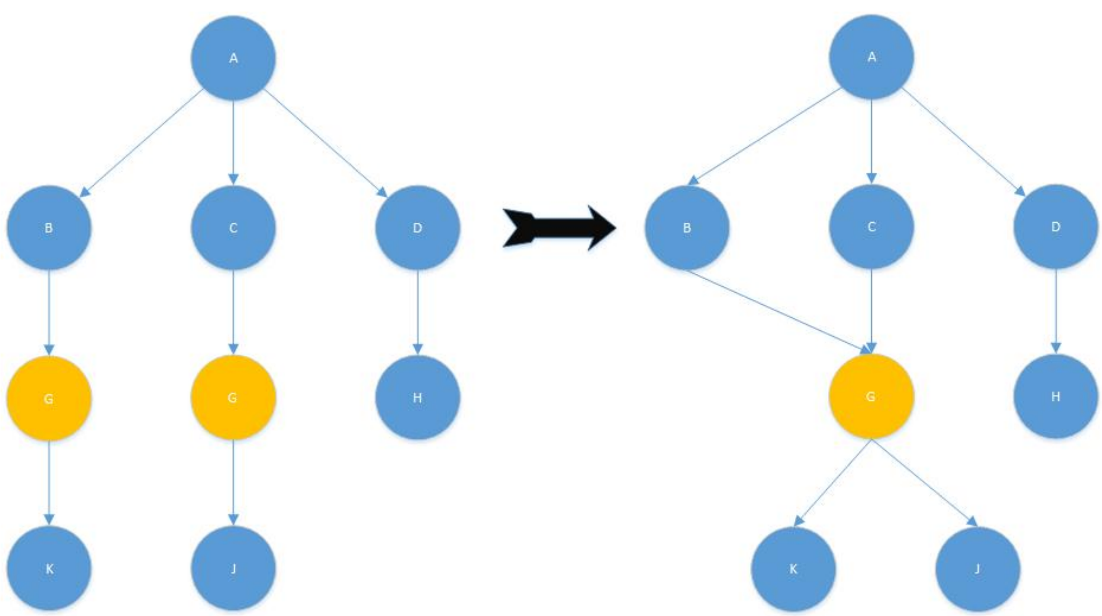

Fig. 9. An example of a decision tree transformation

As can be seen in Fig. 9, nodes B and C had a common child G, with the two nodes G on the same surface. In this case, the two nodes become one and their offspring are added to the new node. Conditions may arise, as in Fig. 10, where the two nodes G are merged, but the situation is that the first node G has a child C and the second node G has a child B and each of these nodes for the opposite G node. It has been visited in the past. This is not a problem either, and it can only be said that when writing a bet it can be assumed that having the terms B and C, along with its logical And, crosses the path G, that is, A and (B OR C). AND G.

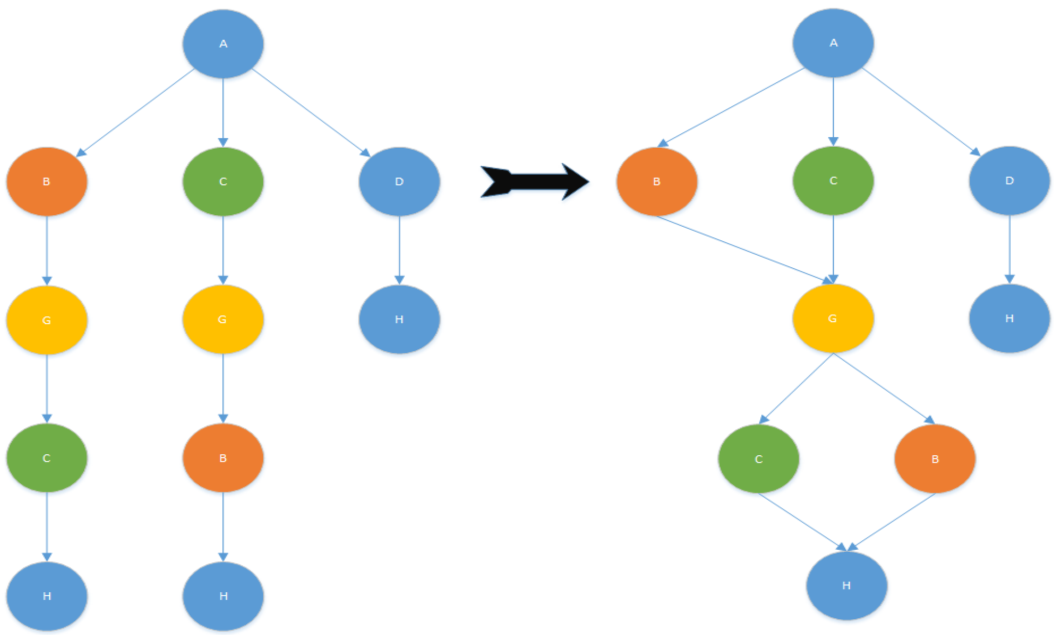

Fig 10. An example of a particular mode of conversion

Here we find that each of these attributes in each data can have a different value, for example in Fig. 11, G can have either true or false values for this part of our data. We choose the one whose frequency is greater in the corresponding attribute in the whole data set. That is, for example, the value of true is greater or false among the G-type attributes then we choose it. If the value of the two values is equal then in the decision we specify or for these two values, for example G = true or G = false.

Extract decisions from the decision graph made here. To do this, start from the root node and make a decision towards each leaf we go to.

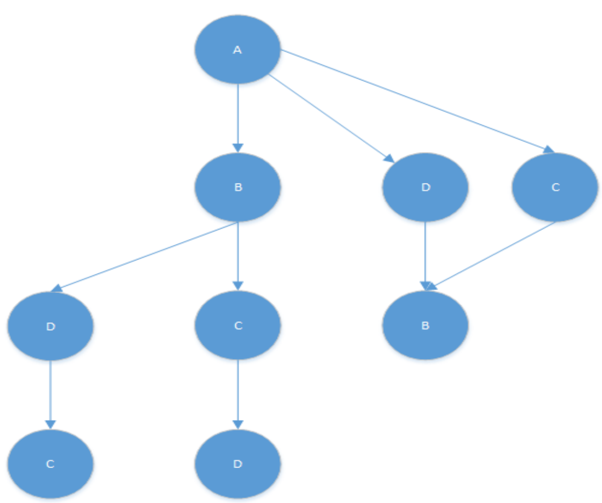

Fig. 11. An example of a decision graph of the proposed method

3.7. Support Vector Machine in Proposed Model

In this step, the normalization process is first performed and then the results of the normalization are extracted, part of this data is used as training data and the vector machine model is created and then the weight test data of this algorithm is created. It is computed until we can then calculate how effective this part of the algorithm is.

4. Results and Evaluation of the Proposed Model

In this section, the proposed method described in the previous section is examined and the proposed method is compared with the known algorithms called ID3, SVM algorithm and Bayesian network. As mentioned in the previous section, the proposed solution is based on ID3, which has more advantages than the ID3 method, and in this section the performance improvement of the proposed method is observed. The proposed approach is evaluated using a three-year data set of companies that indicates whether or not the company is fraudulent.

The data is first formatted by the program in the appropriate format for analysis, or pre-processed, ie, the ARFF file is created, which is a standard structure suitable for analysis. Implemented in Visual Studio 2017 with C # programming language and help from libraries like za and zedgraph while working.

The system used here is Windows 10, has 6GB of RAM and Corei7. In this study, the proposed algorithm is compared with ID3, SVM and Bayesian algorithms. The ID3 and SVM algorithms are the basic algorithms of the proposed method which is derived from the combination of these two methods and is also compared with one of the famous algorithms called the Bayesian network which can be followed by the results. Observe the buildup for these algorithms.

```
--------Naive Baysian----------
confusionMatrix:
[0,0] = 18 [0,1]=11
[1,0] = 27 [1,1]=124
Correct Prediction Percent = 78.8888888888889%
InCorrect  Prediction Percent = 21.1111111111111%
MeanAbsoluteError(MAE) = 0.215093567412967
MeanSquaredError(MSE) = 0.450450314025597
RelativeAbsoluteError(REA) = 57.1003786873271
Correct Prediction Number = 142
InCorrect Prediction Number = 38
TP: 124
FP: 11
FN: 27
TN: 18
```

Fig 12. The output of the Bayesian algorithm

```
--------MyAlgorithm----------
confusionMatrix:
[0,0] = 18 [0,1]=9
[1,0] = 27 [1,1]=126
Correct Prediction Percent = 80%
InCorrect  Prediction Percent = 20%
MeanAbsoluteError(MAE) = 0.251683833684513
MeanSquaredError(MSE) = 0.390570453145535
RelativeAbsoluteError(REA) = 66.813909805457
Correct Prediction Number = 144
InCorrect Prediction Number = 36
TP: 126
FP: 9
FN: 27
TN: 18
```

Fig 13. The output of the proposed algorithm

```
--------ID3----------
confusionMatrix:
[0,0] = 7 [0,1]=7
[1,0] = 38 [1,1]=128
Correct Prediction Percent = 75%
InCorrect  Prediction Percent = 25%
MeanAbsoluteError(MAE) = 0.343346480854847
MeanSquaredError(MSE) = 0.430065588743028
RelativeAbsoluteError(REA) = 91.147375133409
Correct Prediction Number = 135
InCorrect Prediction Number = 45
TP: 128
FP: 7
FN: 38
TN: 7
```

Fig 14. The output of the ID3 algorithm

```
--------SVM----------
confusionMatrix:
[0,0] = 2 [0,1]=4
[1,0] = 43 [1,1]=131
Correct Prediction Percent = 73.8888888888889%
InCorrect  Prediction Percent = 26.1111111111111%
MeanAbsoluteError(MAE) = 0.261111111111111
MeanSquaredError(MSE) = 0.510990323891863
RelativeAbsoluteError(REA) = 69.3165467625899
Correct Prediction Number = 133
InCorrect Prediction Number = 47
TP: 131
FP: 4
FN: 43
TN: 2
```

Fig 15. Output related to SVM algorithm

According to the obtained results, it can be clearly seen that the proposed algorithm has the highest accuracy with 80% accuracy and the least error with 20% error. Table 2 compares the proposed algorithm with other algorithms. It is quite clear that the proposed algorithm performs much better than other algorithms, and much better than the ID3 algorithms themselves and the support vector machine, which is the basic algorithm of the proposed method.

Table 2
Ratio of accuracy of predictions and error of predictions

	Percentage of correct predictions	**Percentage of incorrect predictions**
Proposed Ago.	80%	20%
Baysian Ago.	78.88%	21.11%
ID3 Ago.	75%	25%
SVM Ago.	73.88%	26.11%

Given the precision obtained, it is quite evident that the proposed method performs much better than the other methods, while the Bayesian algorithm performs better than the ID3 and SVM algorithms. The ID3 algorithm also performs better than the SVM algorithm. It can be clearly seen here that the proposed hybrid method can achieve higher accuracy than the Bayesian network by integrating ID3 and SVM methods, each of which have lower accuracy than the Bayesian network. The proposal uses improved ID3 and the improvement process outlined earlier. This tree has the lowest possible height and hence has the lowest overhead.

The diagrams obtained from the program are discussed below. The first graph is the percentage of correct prediction among the test data, which can be seen in Fig. 16.

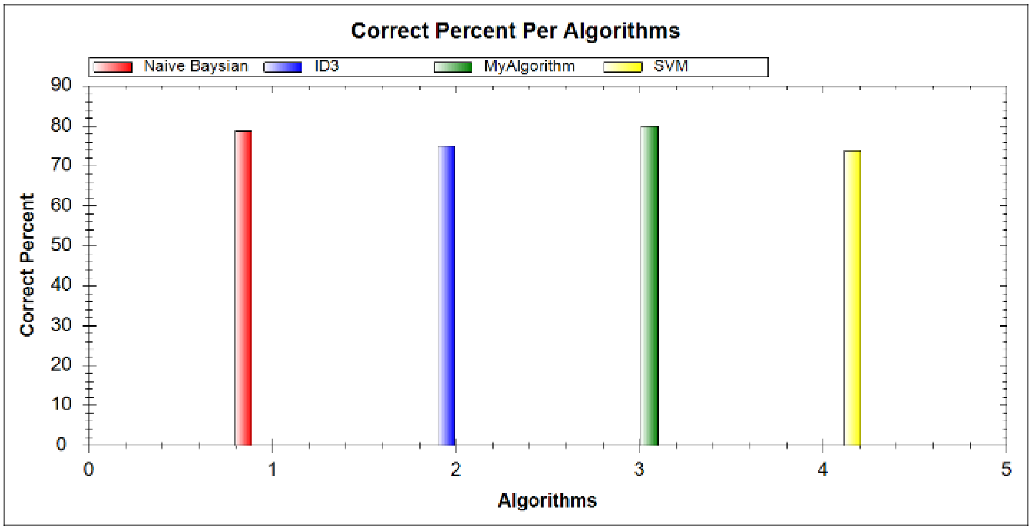

Fig 16. Proportion of correct prediction among test data for the proposed algorithm and other algorithms investigated

As can be seen from this graph, our proposed method has more accurate prediction accuracy than other Bayesian algorithms, SVM and ID3 algorithms. This is because in our prediction method we considered only those test data that had the greatest impact on output and therefore did not use the data that did not affect output, thus greatly reducing the time of analysis. We reduced it while other algorithms are less accurate because of using all the parameters because some parameters may have distances which may have no effect on the output but because in other algorithms we build the model for pre The nose of these parameters have been used to create noise and reduce noise The proposed algorithm also employs a hybrid approach that can be seen to perform well and perform much better than other methods and perform better than the methods based on them.

In the graph presented in Fig. 17 we can see the percentage of incorrect forecasts. According to this graph, it can be understood that the proposed method has lower values than the other methods because it was said earlier with the correct prediction percentage, so the proposed method is less accurate, so the proposed method works better than the other methods has done.

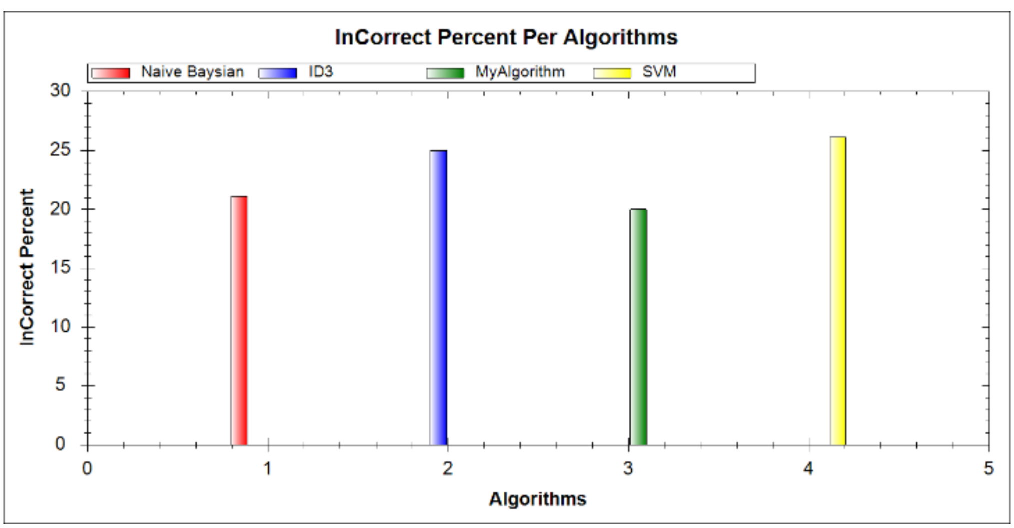

Fig 17. Incorrect prediction percentage among test data for the proposed algorithm and other algorithms investigated

According to the diagram in Fig. 17, it can be concluded that the prediction error is almost equal in the studied algorithms, but the proposed algorithm is lower because the higher the accuracy rate, the lower will be the false rate. And this shows the proper performance of the proposed method. It can be seen that the mean error rate (MSE) in the proposed method is lower than all other methods and the ID3 algorithm is lower than the SVM algorithm, while the Bayesian algorithm has a lower error rate than the SVM algorithm. This error rate not only checks for the incorrectness but also calculates the distance of the predicted answer from the actual answer, in which case the proposed algorithm behaves much better than the other algorithms. If we look at this graph we can see this because the ID3 algorithm has a worse prediction than the Bayesian algorithm, but since the distance was not considered, the ID3 was worse than the Bayesian network but can be seen in Fig. 18 Observe that the ID3 algorithm has a lower MSE than the Bayesian algorithm, which means it has a lower error rate. In general, MSE is very important in data mining and is highly credible.

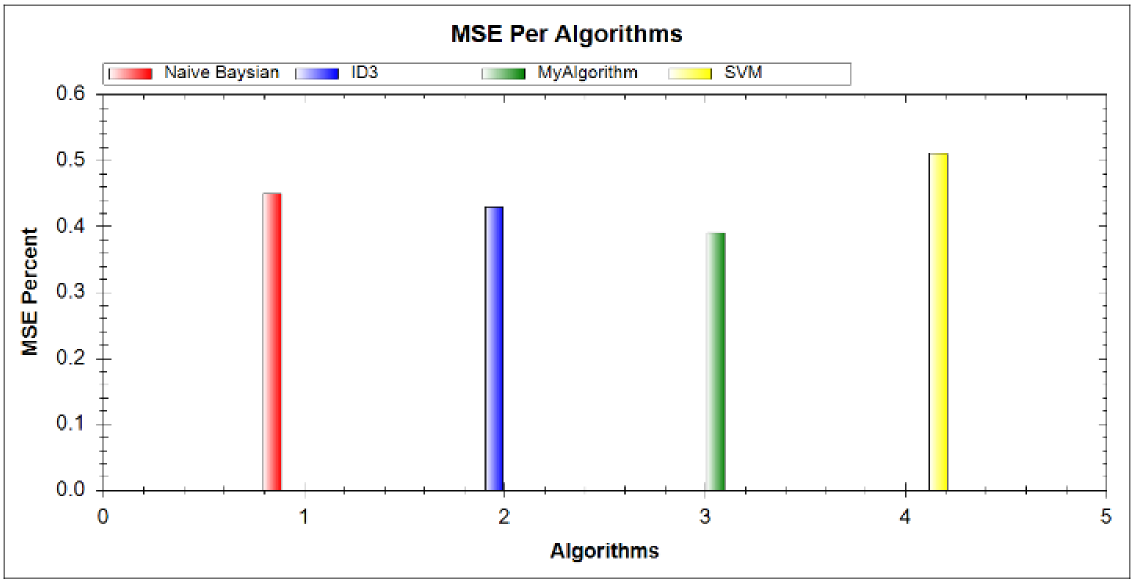

Fig. 18. The MSE benchmark among the proposed algorithm and other similar algorithms is investigated

Following is the Confusion Matrix for the proposed method and other methods investigated in this study.

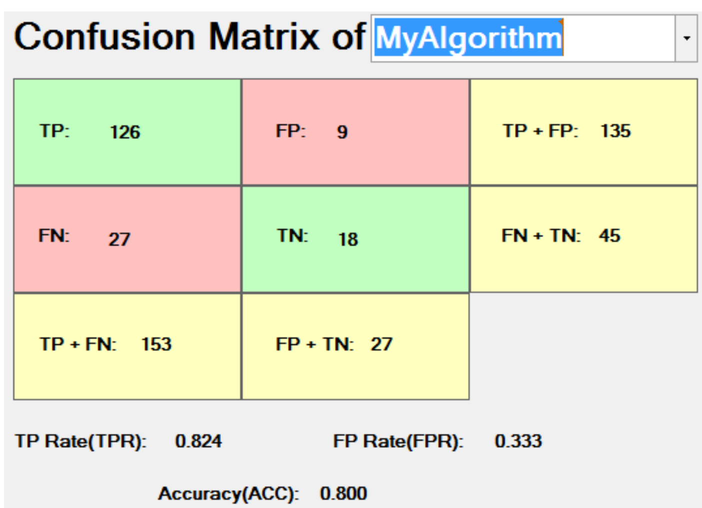

Fig 19. Confusion matrix for Suggested method

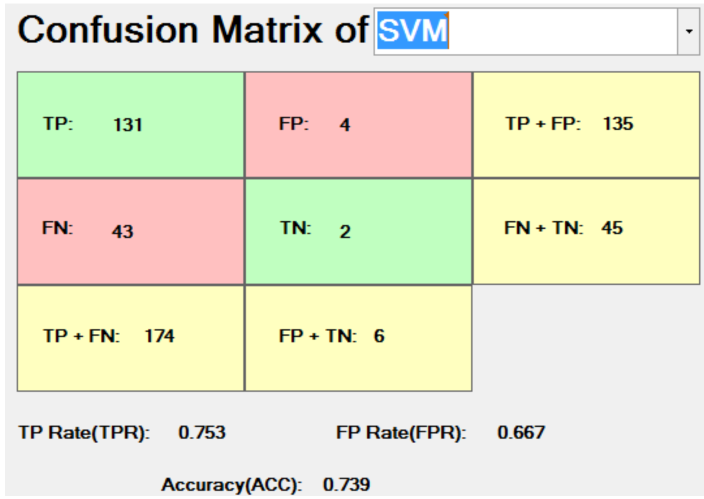

Fig. 20. Confusion matrix for SVM

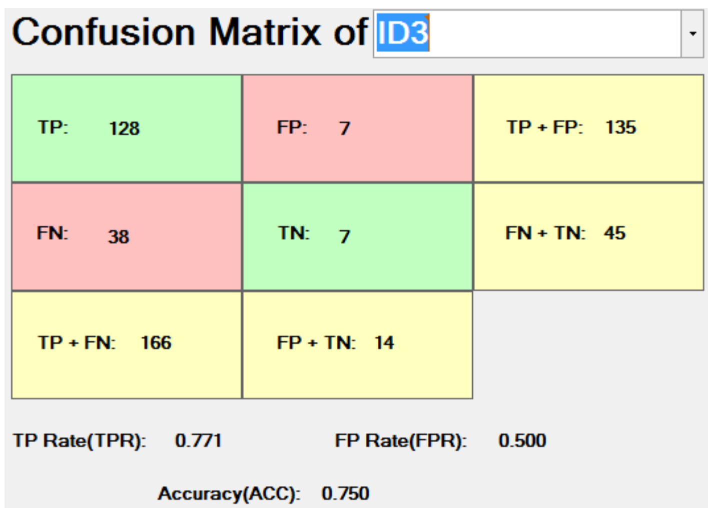

Fig. 21. Confusion matrix for ID3

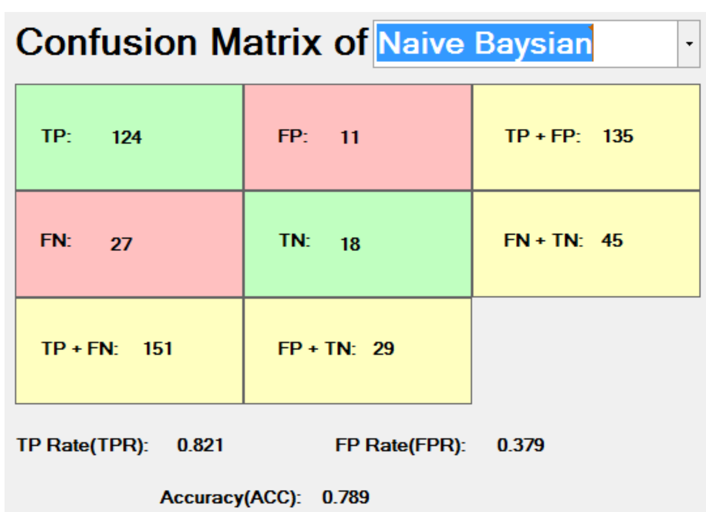

Fig. 22. Confusion matrix for Baysian

Here, it can be seen that the proposed method is more accurate because it contains more TP and TN values than the other algorithms and also has less FP and FN than the other algorithms investigated here. . Because the higher the algorithm, the higher the TP and TN, which means that the test data set is more or less false, and the FP and FN represent the opposite, which is a false prediction.

5. Discussion and Conclusion

Knowledge is nowadays a valuable and strategic resource as well as an asset for evaluation and forecasting, and providing these solutions in the discovery of fraudulent companies increases the accuracy as well as decreases the effective and permanent workforce for fraud detection and detection. That is to say, a solution such as the one proposed can be fully investigated by fraudulent companies, and this does not require a human workforce, but rather the system itself can intelligently detect and provide information. In this study, a strategy was presented to evaluate and predict corporate financial fraud forecasts, and it was found that the method presented here performed well and showed a relatively high improvement over its basic algorithms, ID3 and SVM. Proposed 6.66 percent improvement over ID3 algorithm and 8.27 percent improvement over SVM. Further work with the Bayesian network algorithm was also investigated and it was found that the proposed method performs much better than the Bayesian algorithm and has higher accuracy and lower error rates. The Bayesian algorithm performs much better than the SVM and ID3 algorithms, but it is observed that if the MSE error rate is investigated, the ID3 has a lower error rate than the Bayesian, because the MSE is only dependent on TP, TN, FP and FN is not. The data used in this study were 60 companies over a period of 3 years, which means that the data analyzed here had 180 records. Here, the data was initially processed and transitions were performed on the data until the data became the input data required by the proposed algorithm. The results show a complete improvement of the proposed method. Therefore, with respect to the results of the model presented with 80% accuracy, it has the highest accuracy and 20% error with the lowest error rate, which can be used to predict fraud or as a representative of fraud in various surveys.

6. Suggestions for Future Research

- In the proposed method, entropy is used, but other methods can be used or merged with other methods. For example, if we combine this with a value-based method like Gain, it is likely to perform better because entropy also has disadvantages, but its speed is high, and here we were looking for a method that The speed is high, but the accuracy of the proposed method can be greatly enhanced by integrating this method or alternative methods.
- The proposed method can also be used with the improvement in C4.5 algorithm, ie the proposed method can be implemented on C4.5 with the same improvement as on ID3 in this study, so that its performance cannot be increased but it cannot be Cut said that its performance is improving, but it should be tested to verify the performance.

References

Andon, Pa., Clinton F., % Benjamin S. (2015). "Pathways to accountant fraud: Australian evidence and analysis", *Accounting Research Journal, 28(*1), 10-44, 2015.

Daghmeh Q., Firouzajai, A. (2014). Accounting in Financial Reporting: Disclosure of Fraudulent Companies, MSc, Accounting, University of Mazandaran.

Dionysios S. D. (2018). Fighting money laundering with technology: A case study of Bank X in the UK, Decision Support Systems, 105, 96–107.

Dreżewski, R., Sepielak, J., Filipkowski, W. (2015). The application of social network analysis algorithms in a system supporting money laundering detection, Information Sciences, 295, (20), 18-32.

Etemadi, H. & Zalaghi, H. (2013). Application of Logistic Regression in Identifying Fraudulent Financial Reporting, *Journal of Auditing Knowledge*, 13, (51), 144-163.

Faghandoust Haghigh, K. & Borouari, F. (2009). Investigating the Use of Analysis Methods in Risk Assessment of Financial Statements (Management Fraud), *Journal of Accounting Knowledge and Research*, 16, 18-70.

Farzai, S, Hosseini Shirvani, M., Rabbani M. (2020). Multi-Objective Communication-Aware Optimization for Virtual Machine Placement in Cloud Datacenters, *Sustainable Computing: Informatics and Systems*, doi: https://doi.org/10.1016/j.suscom.2020.100374.

Halbouni, S. S. (2015) The Role of Auditors in Preventing, Detecting, and Reporting Fraud: The Case of the United Arab Emirates (UAE), International Journal of Auditing, 19, 117–130.

Hosseini Shirvani, M. and Babazadeh Gorji, A. (2020). Optimisation of automatic web services composition using genetic algorithm, Int. J. Cloud Computing, 9(2), 1–15.

Iranian Audit Organization (2015 and 2005), Auditor's Responsibility for Fraud and Error in Auditing Financial Statements, Auditing Standard 240, Tehran.

Kantesh Kumar O., Xu De. Z. & Pinial Khan B. (2014). A Fuzzy Rule based Approach to Predict Risk Level of Heart Disease". Global Journal of Computer Science and Technology: C Software & Data Engineering, 14(3) Version 1.0

Kim, Yeonkook J., Baik, B. & Cho, S. (2016). Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning, Expert Systems with Applications, 62, 32-43.

Lari, Dashtbayaz, M. (2015). Data search and discovery process for financial statement fraud, *Research Journal of Finance and Accounting*, 6, (3).

Lookman, S., & Selmin N. (2015). "A Framework for Occupational Fraud Detection by Social Network Analysis", In CAISE FORUM.

Maham, K. & Torabi, Abolfazl (2012), Presentation of Risk Rating Model in Financial Reporting Fraud, Economic Jihad Conference (Emphasizing on National Production, Support for Iranian Labor and Capital), University of Mazandaran.

Mohamed Yusof. K., Ahmad Khair A.H. & Jon Simon., (2015). Fraudulent Financial Reporting: An Application of Fraud Models to Malaysian Public Listed Companies, The Macrotheme Review. 4(3)

Ojeme Blessing, O. & et. al, (2014) "Fuzzy Expert System for Malaria Diagnosis"., An International Open Free Access, Peer Reviewed Research Journal, Published By: Oriental Scientific Publishing Co., India, 7,(2), 273-284 [ISSN: 0974-6471].

Rezaee, Z., and R. Riley, (2010). Financial statement fraudprevention and detection. 2nd edition, John Wiley & Sons, Inc

Senthil Kumar, A.V. & et. al. (2013). Diagnosis of heart disease using Advanced Fuzzy resolution Mechanism*" International Journal of Science and Applied Information Technology (IJSAIT)*,2 , (2), 22-30.

Sergio, L., Gustavo C., Gabriel, A., Donato H., & Miguel T. (2008) "Feature Selection Using Artificial Neural Networks", A. Gelbukh and E.F. Morales (Eds.): MICAI 2008, LNAI 5317, . 351–359,

Sudan C. (2016), Detection of fraudulent financial statements using the hybrid data mining approach, Springer Plus, 5, 89.

Vakili fard, H. & Ahmadi, A. (2010), Investigating the Characteristics of Fraud in the Financial Statements, Journal of Accounting, 210, 36-41.

Ziming Yin, Yinhong Zhao, Xudong Lu, and Huilong Duan, (2014), Screening of Alzheimer's Disease Based on Multiple Neuropsychological Rating Scales, Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine, Volume 2015, Article ID 258761, 13 pages.

Online References

https://chistio.ir (2019)

http://farsithesis.ht3.ir/farsithesis/41484/html (2018)

http://hkamal.persiangig.com/document/genetic.Doc (2018)

Javadian Kootanaee, A., Poor Aghajan, A., Hosseini Shirvani, M. (2021). A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements.
Journal of Optimization in Industrial Engineering, 14(2), 183-201.

http://www.qjie.ir/article_674755.html
DOI: 10.22094/JOIE.2020.1877455.1685