A Data Mining Approach To Detection Financial Distress In Iraqi Companies

Dalya Abdulkarim Abdullah[a], Nashaat Jasim AL-Anber[b]
[a]Information Technology Department, Technical College of Management-Baghdad, Middle Technical University,Iraq. Email: dalyakarim1994@gmail.com
[b]Assistant Prof. Information Technology Department, Technical College of Management-Baghdad, Middle Technical University,Baghdad, Iraq. Email:dr.nashaat@mtu.edu.iq
Article History: Do not touch during review process(xxxx)

Abstract: Due to the difficulties experienced by the financial auditors and the management analyst, in order to know the financial performance of the company and the ability of companies to continue and because of the inconsistency of the financial information being not transparent so renewed the direction of accounting work to use artificial intelligence methods and data mining techniques. In this paper, data Mining (DM) and deep learning (DL) methods were used to detect financial distress, using Artificial Neural Networks (ANN) algorithm represented by the Multilayer Perception Feed Forward Neural Network Error Back Propagation Algorithm (MLP-FFNN) as well as the C4.5 algorithm and the Multi-class support vector machine (MSVM).The results of the analysis showed that the C4.5, ANN and MSVM algorithm had the highest rate of rating accuracy by a small margin on all scales and were (97.98 , 96.97 , 91.92) respectively. In this study, the data of companies listed on the Iraq stock exchange for 2017 were taken, including 36 companies with high financial distress, 20 with medium financial distress and 43 non-distressed for a group of 99 companies .
Keywords: Financial Distress, Financial ratios, C4.5, Support vector machine, Artificial Neural Networks, Multilayer perceptron

1. Introduction

The stock market has become an important source of economic growth and domestic and international finance because it is the tool through which the economic units are fed as Iraq is one of many countries that seek to develop and modernize its financial market to make it more modern to keep pace with the technological financial development, so the phenomenon of financial distress is one of the Appropriate to protect shareholders and stakeholder ,The identification of financial distress indicators is the cornerstone in ensuring the achievement of the most important objectives that the company seeks, namely survival, growth, and continuity, survival, growth and continuity are related to the financial situation of the company and its achievement of profits and its ability to cope with future crises as well as the ability to overcome these crises .The economic sector is one of the important sectors that work to attract investments in order to contribute to the achievement of sustainable development, The integration of the advantages of information technology with statistical methods and algorithms has led to the availability of the necessary possibilities to predict future behavior and then to develop appropriate solutions to problems before they occur if they can occur, or from the side of forecasting with the aim of development and modernization in general in various fields, all using data mining techniques[1][2]. The main objective of financial distress

detection models is to determine whether the company will be exposed to financial distress in the future as well as to detect bankruptcy and insolvency, which are other aspects of financial distress that discriminatory analysis and statistical analysis are among the initial and traditional models used in the field of financial distress detection these techniques are traditional linear and It's unrealistic and therefore can't be used to generate a powerful predictive model. In 2008, Park conducted a study that focused on the extent of ANN's ability to predict the bankruptcy of the sample of companies operating in the hospitality field in the United States of America compared to other classical statistical methods, in addition to the importance of knowing the best ratios used to distinguish between companies exposed to and not exposed to bankruptcy, for a sample of 128 companies with 18 ratios and financial index, The study found that ANN models are able to predict the financial impact of companies with excellent accuracy and have an advantage over classical statistical models[21], In 2009, researchers Aghaie and Saeedi developed a model for forecasting the financial stumbling of Tehran Stock Exchange Companies for a sample of 72 distressed companies and 70 non-stumbling companies from 1997 to 2007, using the NB model, (LR) based on 20 financial indicators, and the researchers found that the accuracy of the prediction using the NB model was 90% equal to the accuracy of the prediction (LR) built for comparison with the NB model[22]. In 2012 Erdogan applied the SVM method to analyze bank bankruptcy by determining important financial ratios, the researcher used the data set from Turkish commercial banks and found that the SVM method is able to extract useful information from the financial statements data and can be used as part of the early warning system to predict financial bankruptcy [23]. In 2018 Ruxanda et al researchers built a model to classify companies on the Bucharest Stock Exchange into binary (low and high) tripping classifications using classification methods(Decision tree, support vector machine, logistic regression, fisher linear classifier), a large number of financial ratios were used so the method of Principal component analysis (PCA) was applied to choose the most important financial indicators to predict accuracy rates up to 90% in the training sample and 87% in the test sample[24]. In the current year 2021 researcher Chyan long jan built high-precision and effective forecasting models to predict financial stumbling through the use of deep learning algorithms represented by (ANN), (CNN) important variables were also selected by (Chi-Squared), and the data of Taiwan sample companies from (Taiwan Economic) Database from 2000 to 2019 for(80) distressed companies and(258) non-distressed forecasting financial stumble of 94.23% and the lowest error rate of Type I and error rate of Type II which were 96% and 4.81% respectively[25].

2.Research method

1. Data collection

The data used in this study obtained from Iraq stock exchange. Based on the background of Iraqi listed company for (99) companies, companies have been classified into three classes (categorical) based on their financial condition (non-distressed, medium financial distressed, high financial distressed) for one year (2017) including 36 companies with high financial distressed, 20 with medium financial distressed and 43 non-distressed.

2. Feature Selection

Financial Ratio is the most important tool on which management relies in analyzing financial statements to determine the safety of the financial situation and profitability of the company, as it is relied on by other parties, especially owners and lenders when making investment decisions, as well as to improve the company's future performance, financial ratios (financial indicators) are defined as a tool to extrapolate the, The first stage of this process was to manually extract the financial statements (income list, Financial Position list, statement of shareholders ' rights) from the annual report of the research sample companies, and in the second stage the financial ratios were calculated based on the financial statements, and from this point the researcher came to choose (14) financial ratio its components, if We can use them to build a model that is more accurate in classification and forecasting, taking into account its suitability for the Iraqi corporate environment.

Table.1 Definition of classification variables

Variable	Financial ratios	Variable	Financial ratios
X1	Current assets to current liabilities	X8	Total liabilities to assets
X2	Current liabilities to total assets	X9	Property rights to total assets
X3	Current assets-current liabilities to total assets	X10	Shareholders ' equity to total liabilities
X4	Income to total assets	X11	Gross profit to assets
X5	Revenue to equity	X12	Net profit to net income
X6	Income to cash	X13	Total assets to liabilities
X7	Retained earnings to total assets	X14	Cash to total assets

3.Technical background

1. Support vector machine

(SVM) A modern educational system designed by the scientist Vapnik in 1992 based on statistical learning Theory, is one of the supervised learning algorithms used in classification and regression, and is often used in solving classification issues for its effectiveness and for obtaining excellent accuracy in most types of data used where it is a method of classification of both linear and nonlinear data, (SVM) algorithm possesses processing, that the main idea of the SVM algorithm is to find the best Super level (Optimal Hyperplane) which divides the totals (Classes) in the best possible way, the classification process is done by finding the hyperplane and defining it to distinguish between two categories, it is noted that the farther the points from the Super level the greater the achievement of high and correct classification accuracy, when starting to train the algorithm will have more than one hyperplane in this points from data sets, In this case we have to find the best hyperplane by choosing the hyperplane in which the margin is between it and the data points is large points that ara close to the hyperplane ,The margin represents the distance between the hyperplane and the nearest point of the datasets this point called Support vectors and these points if removed from the data set will change the location of the hyperplane that divides the data, so these points are important elements of the data set [3,4,5,6].

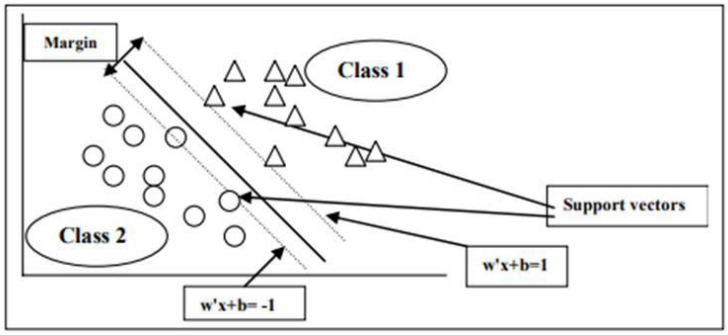

Figure .1 The Hyperplane and Support vectors and margin [7]

The Hyperplane used to separate the data is determined by the following equation:

$$w.x + b = 0 \quad (1)$$

Where

w: Weight vector

b: Bias

x: Represent attribute values

Where the value of x is compensated by the equation according to the following:

$$w.x + b \geq 1 \quad (2)$$

When X belongs to the positive category

$$w.x + b \leq -1 \quad (3)$$

When X belongs to the negative category

Where $w.x + b = 1$ The hyperplane of the positive category will be obtained

As for whether $w.x + b = -1$ the hyperplane of the negative category will be obtained.

So the maximum margin is calculated by selecting the minimum (W) through the following equations:

$$D1 = w.x + b = 1 \qquad w.x + b - 1 = 0 \tag{4}$$

$$D2 = w.x + b = -1 \qquad w.x + b + 1 = 0 \tag{5}$$

By algebraically solving we get the following:

$$w.x + b - 1 - w.x + b + 1 \tag{6}$$

$$m = max\ \frac{2}{/w/} \tag{7}$$

Where:

m : Represents margin for hyperplane optimal

Or maximum margin can be obtained from the following equation:

$$m = min\frac{1}{2} / w /^2 \tag{8}$$

There are two types of margin, the first type is called hard margin or Linear SVM, the data set is linearly separable, i.e. the data points are not overlapping with each other [4]

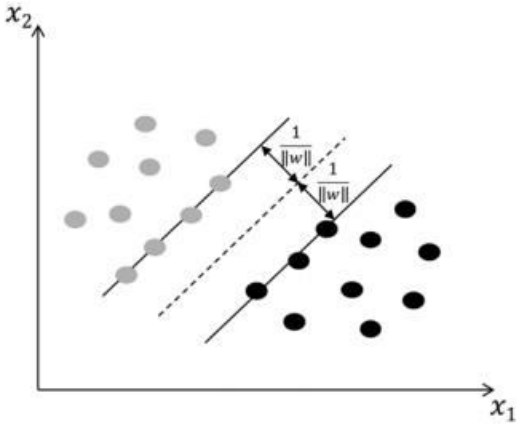

Figure.2 Hard Margin [4]

The second type is called soft margin or non-Linear SVM, the data set contains some points that prevent the process of separating the data linearly, in which case these points are allowed to be incorrectly classified, so a coefficient is added that expresses the percentage of error at each training point whether it is correctly or incorrectly classified.

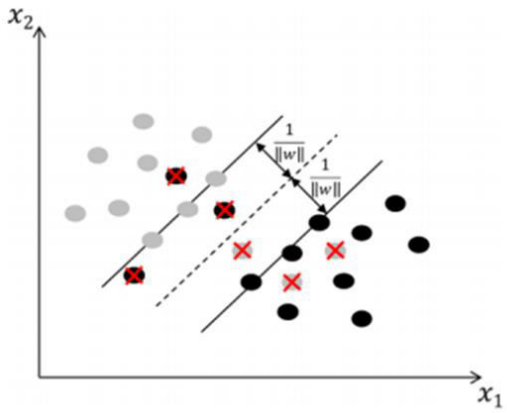

Figure.3 Soft Margin [4]

2. Multi-Class Support vector machine

SVM has been developed to work with multiple categories, which are called Multiclass support Vector machine Algorithm, as many MSVM methods analyze training data into several binary categories based on a number of methods, including one against one (OAO) method and one against all (OAA) and other methods.

- One Against One: A basic method that enables us to create binary classifiers toseparate each pair of classes where Class 1 is vs. Class 2, Class 1 is vs. Class 3 and Class N-1 is vs. Class N, this method will build a binary classifier of the size of N (N-1)/2 after which new objects will be classified for the class that.
- One Against All :The method (OAA) is one of the simplest and fastest methods (MSVM) where the existence of N of the categories we want to classify, we create N of binary classifications to separate each category from the rest of the categories, meaning that category 1 is positive versus that all other categories are negative, and new objects are assigned to the category that has a positive [4] [5].

3. C4.5 Algorithm

The C4.5 algorithm was introduced for the first time in 1993 by Ross Quinlan, which was considered as an extension of previous studies that led to the development and improvement of an earlier algorithm called ID3 developed between 1970 and 1980, that the C4.5 algorithm is one of the types of algorithms that work in the method of Decision Tree works to classify cases C4.5 algorithm under the group of learning algorithms operating under the supervision of supervised which The complex issues are divided into simpler ones, and then the function itself and all parts of the issue are self-invoked, and by combining the solutions of the fragmented issue, the solution of the complex issue is reached,C4.5 has become a standard for evaluating and comparing other classification methods, C4.5 uses a feature test metric called The Information Gain scale based on a qualitative scale called Shannon Entropy that reflects the randomness or purity of the data set [8][9].

The work steps of the C4.5 algorithm are as follows:

Let D be the complete training dataset and N are the class Label titles associated with it, and let C_j be the list of descriptors that include m from classes $(C_{1,}C_{2,\ldots\ldots,}C_m)$ and let $C_{j,D}$ be a subset of the groups that belong to C_j in the Training Group D, and let $|C_{j,D}|$ and $|D|$ represent the number of Tuples of both $C_{j,D}$ and D respectively.

The Entropy scale of the data set is represented by the following equation:

$$Entropy(D) = -\sum_{j=1}^{m} p_j \log_2(p_j) \quad (9)$$

Where:

$$p_j = |C_{j,D}| / |D| \quad (10)$$

p_j : Represents the probability that a particular group belongs to the category C_j

Entropy measures the homogeneity or heterogeneity of a particular set of data based on the designations of the category concerned, where the data set is said to be pure or homogeneous if all its sets belong to one taxonomic category, and if the data sets are different in the taxonomic category the data set is considered to be Impure or heterogeneous.

Suppose a dataset D is divided into random classes (attributes) A that includes v of the characteristic values,$\{a_1, a_2, \ldots\ldots, a_v\}$, and D can be divided into subgroups v $\{D_1, D_2, \ldots, D_v\}$ and each subgroup includes a set of classes A that have a_j value, and to determine the best division property of the N node in the tree, the ID3 algorithm used a division standard called the information Gain standard.

$$Gain(D, A) = Entropy(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Entropy(D_j) \quad (11)$$

The information Gain measure represents the difference between the original entropy representing the minimum information required to classify a given group in D and the expected value of entropy after dividing D by attributes a, the class with the highest information gain (minimum entropy) is chosen as the division attribute in Node N. The C4.5 algorithm also uses another division measure known as Gain Ratio, which can be represented by the following equation:

$$Gain\ Ratio(D, A) = \frac{Gain(D,A)}{-\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2(\frac{|D_j|}{|D|})} \quad (12)$$

4. Artificial Neural Networks Models

(ANN) is a mathematical software technology designed to simulate the work of the human mind (the biological neural network of the human being), by performing extensive processing distributed in parallel by simple processing units called neurons (Neurons) or nodes (Node) each connection between these cells is defined by values called Weights, ANN is fast and simple as it acquires knowledge through its ability to learn and train and stores this knowledge in the neuron to make it available to the user, through the process of adjusting weights [10][11][12].

4.1 Structure of artificial neural networks

The structure of the artificial neural network (ANNs) consists of the following parts [13] [14] [15].

1. Input layer: receives input signals from outside the network to feed the network the required variables, as this layer is limited to transferring data to the processing units through the hidden layer.

2. Weights: relative coefficients within the network determine the intensity of the input signal as recorded by the artificial neuron, each input has a relative weight that gives the input the effect it needs on the function of collecting processing elements, weights are a measure of the strength of the input connection, these weights can be adjusted in response to different training groups and network learning rules.

3. Hidden layer: this layer receives the data sent from the input layer and then sends it to the output layer via the interfaces after the processing procedure, the choice of the number of neurons in the hidden layer is very important, if the number of neurons in the hidden layer is small, the artificial neural network cant receive all the necessary information and is likely to:

- Number of input and Output
- Units Number of training cases
- Complexity and type of classification problem

4. Summation function: represents the first step in the process of the processing element that calculates the weighted sum of all inputs after multiplying all input elements by their accompanying weights, described by the following equation [16]:

$$y = \sum_i^n p_i w_{i\ j} + b_j \qquad (13)$$

Where in:

y_i : Represents network outputs

p_i : Represents the number of inputs to the system

w_{ij}: Represents the accompanying relative weight of each entrance

b_i: Represents bias value

5.Output layer: the last layer that gives real output after a series of processors performed during the previous layers.

6.Threshold limit: is the limit that determines the type and extent of output, to be compared with the desired output (Target Output) and in other words represents the minimum value at which the input must be to activate nerve cells. Figure (1) illustrates the structure of artificial neural networks.

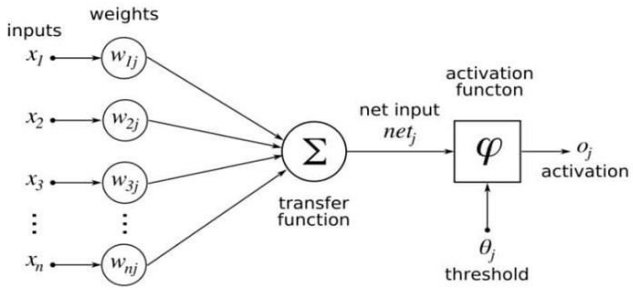

Figure.4 illustrates the structure of artificial neural networks [17]

7. Activation function: The activation function in artificial neural networks represents the extract of the work done in the cell, the activation function is selected depending on the number of layers present in the cell, whenever the number of layers and the number of neurons change the choice of the activation function, also depends on the type of output required, if the values are negative or positive, that each neuron has the relationship that is represented by using the activation function, there are many functions Activation, but there's a few of them have been used in scientific applications, and more de-activation of common use in the field (Classification) is a function Sigmoid Function which be of two types [18][19]:

- Log-sigmoid function: it is one of the most used activation functions in the nodes of hidden layers, as its input has actual values and its output for each node is limited to (0,1), and in most cases the same activation function is used in all layers of the network, represented by the following mathematical formula[8] [11] [13]: $f(x) = \frac{1}{1+exp^{-x}}$ (14)
- The hyperbolic Tangent function is similar to the logistic function but differs in the type of its output where it is limited to(1, -1), and is represented by the formula The following:

$$f(x) = \frac{1-exp^{-2x}}{1+exp^{-2x}} \quad (15)$$

4.2 Error Back Propagation Algorithm

To teach the algorithm BPN by repeating the education (Learning Epoch), you probably will need manyIterations (Epoch) to learn an entire network so you can deal with all the data that have been processed and the end result will be satisfactory, repeat the training process (Learning Epoch) described below, for each entrance to enter in the training data [13][14][16][20]:

1. Feed forward data input
2. initialize weights
3. check the output against the desired value (Actual) and by error as the reverse propagation process contains:
    - calculation of error gradients
    - update weights

In this model, a neural network of the feed forward type was built and a learning algorithm of the back Propagation type was used. The features were entered by a number of (14) attributes from the database for the purpose of network training. Each attribute represents a node of the input layer. the outputs are represented by one node in the output layer, on the basis of which companies are classified into highly stumped, medium stumped, non-stumped. The activation function used in this type of neural network is sigmoid. Figure.5 illustrates the process of training the artificial neural network of the type of reverse propagation network of error in the MATLAB program of the proposed model that was reached after several attempts at training by changing the number of repetitions and changing the number of hidden nodes and layers.

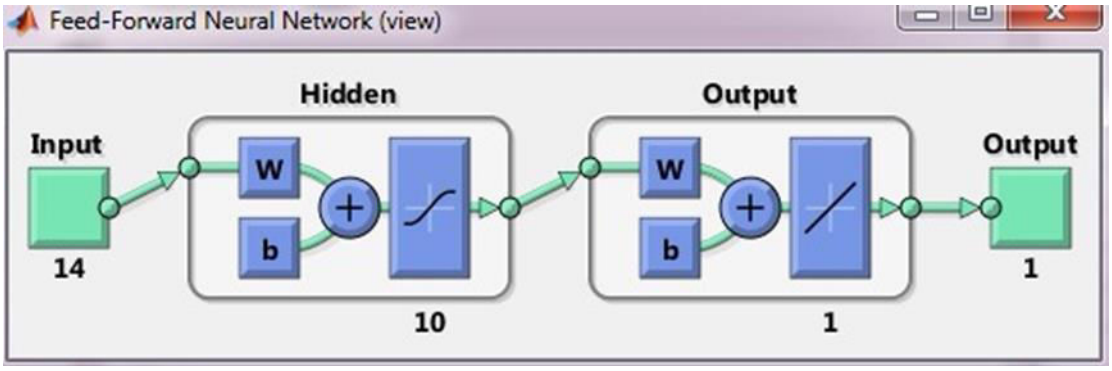

Figure.5 training of Error Back Propagation Algorithm

5. Results and analysis

Classification algorithms (SVM, C4.5, ANN) have shown different and varying percentages of accuracy and predictions for the correct positive and negative category we note from table.2 that the highest rating accuracy rate was for the C4. 5 algorithm and equal to 97.98%, which is excellent compared to the algorithms used because of its effective ability to deal with a small data set with an adjustment rate equal to 95%, and that the highest rating rate was for the ANN algorithm 100% Which comes in the second stage in terms of its ability to classify with 96.97% accuracy and is followed in the third stage by the SVM algorithm with an accuracy rate of 91.92%.

Table.2 The comparison of accuracy results and rating and performance metrics of algorithms

Evaluation Criteria	Classifier		
	C4.5	ANN	SVM
Accuracy (%)	97.98	96.97	91.92
Sensitivity (%)	100	93	93.2
Specificity (%)	96	100	90.9
Precision (%)	95	100	89.13
F-measure	0.97	0.96	0.92
Youden index	0.96	0.93	0.84
BCR	0.98	0.97	0.92

The ideal classifier is often chosen based on the results of the classifier who has a high percentage in the accuracy of the scales, since the F-measure is a consensus measure between the tuning ratio and the sensitivity ratio (recall) and the field values range from (0-1) as the classifier is good the closer the result of each Measure (F-measure, Youden index, BCR) to one, so it was concluded from this study that the best model is C4.5 depending on the results obtained comes after the ANN model with a very small difference.

6. Summary

To the best of our Knowledge, This paper is the first to model financial distress using (ANN, C4.5, MSVM) in Iraqi companies. We show the flexibly of the proposed measure with noise rejection capability and Artificial neural networks reduce input weights and dynamically organize them to improve output, making the application of artificial neural networks widely effective in solving real-world problems.. A series of experiments were conducted in data mining and machine learning in order to detect financial stumbling in companies listed on the Iraq Stock Exchange, which could help the market prosper, and thus serve as an early detection of risks that will lead to the disruption of the functioning of these companies, The results of these experiments were explained and discussed to provide some of the informational knowledge that would perform better for the labor market.

7. Recommendations

1. Government authorities must strictly monitor the financial conditions of companies and enact strict regulations on the financial supervision and governance of companies in order to protect the interests of stakeholders.

2. Encourage and train auditors to practice and use prospecting techniques and smart technologies as they facilitate and accelerate decision-making at the lowest possible cost

3. Research more broadly on the most important financial indicators that have a direct impact on the financial condition of the company and coincide with raising the accuracy of the smart models used , and the need not to rely on the weights of the financial ratios for non-Iraqi studies models because it is prepared according to an economic environment and conditions different from the environment of Iraqi companies.

4. The importance of accessing hybrid algorithms and fuzzy programming for better performance and comparing results with different algorithms applied.

8. References

[1] R. M. Hadi, S. H. J. Al-khalisy, and N. Abd Hamza, "Prediction Model for Financial Distress Using Proposed Data Mining Approach," Journal of Al-Qadisiyah for computer science and mathematics, vol. 11, pp. 37-44, 2019.

[2] K. I. Moin and D. Q. B. Ahmed, "Use of data mining in banking," International Journal of Engineering Research and Applications, vol. 2, pp. 738-742, 2012.

[3] B. Erdogan and E. Egrioglu, " Support vector machines vs multiplicative neuron model neural network in prediction of bank failures," Am. J. Intell. Syst, vol.7, pp. 125-131, 2017 .

[4] R. C. Gilbert, T. B. Trafalis, and I. Adrianto, "Support Vector Machines for Classification," Wiley Encyclopedia of Operations Research and Management Science, 2010.

[5] G. W. Naji and J. M. Al-Tuwaijari, "Satellite Images Scene Classification Based Support Vector Machines and K-Nearest Neighbor," Diyala Journal For Pure Science, vol. 15, 2019.

[6] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction 1," 2011.

[7] Z. Y Algamal,"Support Vector Machine Procedure as a Data Mining Multi-class Classifier," IRAQI JOURNAL OF STATISTICAL SCIENCES, vol.12pp.26-40, 2012.

[8] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, pp. 81-106, 1986.

[9] M. Kantardzic, Data mining: concepts, models, methods, and algorithms: John Wiley & Sons, 2011.

[10] A. S. Mohammad, "Weather temperature forecasting using artificial neural network," Journal of Engineering and Development, vol. 15, pp. 130-139, 2011.

[11] C. A. Arslan, "Rainfall–runoff modeling based on artificial neural networks (ANNs)," European Journal of Scientific Research, vol. 65, pp. 490-506, 2011.

[12] T. Koskela, Neural network methods in analysing and modelling time varying processes: Helsinki University of Technology, 2003.

[13] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques 3rd edition," The Morgan Kaufmann Series in Data Management Systems, 2014.

[14] K. Gurney, An introduction to neural networks: CRC press, 2018.

[15] A. Berrais, "Artificial neural networks in structural engineering: concept and applications," Engineering Sciences, vol. 12, 1999.

[16] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques second edition," The Morgan Kaufmann Series in Data Management Systems, vol. 5, pp. 83-124, 2012.

[17] A. P. D. N. J. Mohammed and H. T. Jaffar, "Build a system for forecasting the electrical load demand in Baghdad," JOURNAL OF MADENAT ALELEM COLLEGE, vol. 9, 2017.

[18] H. N. Mhaskar and C. A. Micchelli, "How to choose an activation function," in Advances in Neural Information Processing Systems, 1994, pp. 319-326.

[19] K. Debes,A.Koenig,and H.-M. Gross, "Transfer Functions in Artificial Neural Networks A Simulation-Based Tutorial," Brains, Minds and Media, vol. 2005.

[20] G. P. Zhang, "Neural networks for data mining," in Data mining and knowledge discovery handbook, Springer, 2009, pp. 419-444.

[21] S.-S. Park, A comparative study of Log it and artificial neural networks in predicting bankruptcy in the hospitality industry: Oklahoma State University, 2008.

[22] A. Aghaie and A. Saeedi, "Using bayesian networks for bankruptcy prediction: Empirical evidence from iranian companies," in 2009 International Conference on Information Management and Engineering, 2009, pp. 450-455.

[23] B. E. Erdogan, "Prediction of bankruptcy using support vector machines: an application to bank bankruptcy," Journal of Statistical Computation and Simulation, vol. 83, pp. 1543-1555, 2013.

[24] G. Ruxanda, C. Zamfir,and A.Muraru, "Predicting financial distress for Romanian companies," Technological and Economic Development of Economy, vol. 24, pp. 2318-2337, 2018.

[25] C.-l. Jan, "Financial Information Asymmetry: Using Deep Learning Algorithms to Predict Financial Distress," Symmetry, vol. 13, p. 443, 2021.