AN OPTIMIZED DEEP NEURAL NETWORK-BASED FINANCIAL STATEMENT FRAUD DETECTION IN TEXT MINING

Ajit Kr. Singh Yadav
Assistant Professor, Department of Computer Science and Engineering,
NERIST, Itanagar, Arunachal Pradesh (India), and Research Scholar,
Department of Computer Science and Engineering,
Rajiv Gandhi University, Itanagar, Arunachal Pradesh, (India).
E-mail: ajityadav101@rediffmail.com ORCID: https://orcid.org/0000-0002-2208-0828

Marpe Sora
Associate Professor. Department of Computer Science and Engineering,
Rajiv Gandhi University, Itanagar, Arunachal Pradesh, (India).
E-mail: marpe.sora@rgu.ac.in ORCID: https://orcid.org/0000-0003-0159-5416

Recepción: 23/07/2021 **Aceptación:** 03/11/2021 **Publicación:** 24/11/2021

Citación sugerida:
Singh, A. K., y Sora, M. (2021). An optimized deep neural network-based financial statement fraud detection in text mining. *3C Empresa. Investigación y pensamiento crítico, 10*(4), 77-105. https://doi.org/10.17993/3cemp.2021.100448.77-105

ABSTRACT

Identifying Financial Statement Fraud (FSF) events is very crucial in text mining. The researcher's community is mostly utilized the data mining method for detecting FSF. In this direction, mostly the quantitative data has utilized by research i.e. the financial ratio is presented for detecting fraud in financial statements. On the text investigation there is no researches like auditor's remarks present in published reports. For this reason, this paper develops the optimized deep neural network-based FSF detection in the qualitative data present in financial reports. The pre-processing of text is performed initially using filtering, lemmatization, and tokenization. Then, the feature selection is done by the Harris Hawks Optimization (HHO) algorithm. Finally, a Deep Neural Network-Based Deer Hunting Optimization (DNN-DHO) is utilized to identify the fraud or no-fraud report in the financial statements. The developed FSF detection methodology executed in Python environment using financial statement datasets. The output of the developed approach gives high classification accuracy (96%) in comparison to the standard classifiers like DNN, CART, LR, SVM, Bayes, BP-NN, and KNN. Also, it provides better outcomes in all performance metrics.

KEYWORDS

Financial statements, Fraud, Non-fraud, Text mining, Deep neural network, Deer hunting optimization.

https://doi.org/10.17993/3cemp.2021.100448.77-105

1. INTRODUCTION

Financial fraud is a major challenging task for different administrations across industries and in several states as it takes vast destruction to business. Due to financial fraud, billions of dollars are lost every year in the bank of America, for instance approves to pay $16.5 billion for solving the case of financial fraud (Rezaee & Kedia, 2012). Material omissions resultant from an intentional failure to report financial data in accordance with usually acknowledged secretarial ethics are termed as FSF (Dalnial *et al.*, 2014). The companies provide the financial statements that include the textual data in the form of auditors' remarks and expose as records with financial proportions. The qualitative data consist of indicators of fraudulent financial reporting in the form of intentionally located idioms. The agents use the adverbial phrases, selective sentence constructions, and selective adjectives to cover the fraudulent activity (Throckmorton *et al.*, 2015; Song *et al.*, 2014). To identify fraudulent financial fraud, financial statement users and regulators expect external auditors. Financial statements are the organization's elementary documents to reflect its fiscal rank (Kanapickienė & Grundienė, 2015).

A careful analysis of the financial accounts can denote whether the corporation is running efficiently or is in crisis. If the corporation is in crisis, financial accounts can show if the maximum dangerous entity handled by the organization is profit or cash or something different (Perols & Lougee, 2011). In every quarter and every year, most of the organizations are needed to publish their financial statements (Gray & Debreceny, 2014).

FSF can be executed to build stock values or to acquire loans from banks. It may be done to allocate smaller profits to investors. One more feasible reason might be to stay away from the expense of assessments (Manurung & Hardika, 2015). Recently, different organizations are creating usage of fraud financial reports to cover up their real fiscal rank and create self-interested improvements at the expenditure of shareholders. In the detection of FSF, financial ratios are prime elements because they present a pure image of the financial strength of the corporation (Hajek & Henriques, 2017).

https://doi.org/10.17993/3cemp.2021.100448.77-105

The economy of an organization is caused by the illegal task of FSF. In determining capitalizing in a corporation, the investigation of financial reports helps the contributors to the investment market (Omar *et al.*, 2014). The performance of the company provided by the data presented in these statements in terms of fiscal rank to the creditors, shareholders, and auditors.

In worldwide organizations, finding and prevention of FSF have become a significant challenge (Gupta *et al.*, 2012a). In the failure of the prevention process, the detection of fraudulent financial reporting is a challenging issue. Though, the prevention of FSF is a better method (Asare *et al.*, 2015). The interior and exterior auditors have to play a significant task in the discovery and prevention of FSF. But they cannot be said only accountable for the identification and detection of FSF (Gupta *et al.*, 2012b). Study about fraud detection and antecedents is significant since it adds to the sympathetic about fraud. To enhance the auditors' and regulators' capability, it has the potential to identify the fraud either directly or by helping as a basis for future fraud research that does (Ravisankar *et al.*, 2011). Better-quality fraud detection can assist the defrauded organizations, and their workers, investors, and creditors curb costs linked with fraud and also enhance the efficiency of the market. This knowledge is interest to auditors once delivering guarantee about whether financial accounts are free of substantial misstatements affected by fraud (Ngai *et al.*, 2011), mainly during audit planning and client selection.

Several researchers have been analysed the quantitative data for the recognition of false financial reporting (Jan, 2018). Therefore, the text mining technique is utilized to recognize fraud and non-fraud financial reports in the qualitative contents of financial statements (Lin *et al.*, 2015). Text mining is the method of mining significant structured data from unstructured text. It can be utilized for finding the fraud or non-fraud reports and also it can examine the words (Gupta *et al.*, 2012c). At present, extensive data is produced from different sources in the Internet-dependent world. In an unstructured format, a vast amount of data is obtainable. Text mining and data mining methods can permit well decision making for analysing unstructured data (Kumar & Ravi, 2016). Different types of tasks involved in text mining, for example, text summarization, web page classification, sentiment analysis, detection

https://doi.org/10.17993/3cemp.2021.100448.77-105

of plagiarism, malware analysis, classification of the document, detection of a topic, patent analysis, etc. In the financial statements, the textual data is unstructured (Dong, Liao, & Liang, 2016). Before applying any data mining approaches like classification or clustering, the text is must be transformed into structured data because the form of text is shapeless for the discovery of FSF.

This work contributes mainly:

- In finding the solution of the financial report fraud discovery.
- To design the model for identifying the fraudulent and non-fraudulent statement.
- To use optimal feature selection approaches to get high accuracy.
- To model a new hybrid classifier for financial statement fraud discovery.

The remaining work of this paper is shown in following sections: Section two defines the recent works related to this paper. The proposed method to detect the FSF is given in section three, the section four provides the outcomes of the simulation and conclusion and future scope is given in section five.

2. RELATED WORKS

An interpretable fuzzy rule-based system was presented by Hajek (2019) for detecting FSF. The developed fuzzy rule-based detection approach combines the rule extraction and element of feature selection to obtain the granularity and rule complexity. A genetic feature selection method is utilized to eliminate the irrelevant features. A qualified investigation of fuzzy systems was performed with evolutionary fuzzy rule-based schemes and FURIA. The developed system leads both desirable interpretability and good accuracy. The result provides significant effects for auditors and other operators of discovery structures of FSF.

Fraud detection was introduced by Chen *et al.* (2019) for economic reports of business groups. For fraud discovery, this article suggests a methodology in the financial reports of business assemblies. The established technique to improve the welfares of investment for creditors and investors and to lessen the

https://doi.org/10.17993/3cemp.2021.100448.77-105

investment losses and risks. The learning points were obtained by the subsequent stages: (i) construct an effective model for fraud discovery in the financial reports of business assemblies, (ii) different fraud finding methods were applied in the financial reports, and (iii) valuation of the developed system.

A Financial Fraudulent Statements (FFS) detection was developed by Temponeras *et al.* (2019) using the deep dense Artificial Neural Network (ANN). This system reviews the financial statements of multiple companies. A deep dense ANN is derived from the decisions about conceivable accounting fraud. To accurately classify the FFS, the data is obtained from 164 Greek companies. Therefore, the main objective was to test a neural system structure in the forecasting FFS. In the classification FFS task, the developed approach provides superior outcomes than other earlier classifiers in investigating the Greek data.

A CHAID, SVM (Support Vector Machine), and C5.0 were discussed by Chi *et al.* (2019) for FSF detection. Through an active detection scheme, an approach of C5.0, SVM, and CHAID are applied to the discovery of FSF. From the Taiwan Economic Journal (TEJ), the research data is obtained. The source sample contains 28 companies involved in FSF and 84 corporations are not intricate in such frauds on the Taipei Exchange and the Taiwan Stock Exchange amid the investigation time. Before constructing the system, the paper chooses key variables with C5.0 and SVM. For FSF, the non-financial and financial variables are utilized to improve the precision of recognition.

An application of a cooperative Random Forest (RF) classifier was presented by Patel *et al.* (2019) for identifying financial report management of Indian registered corporations. Recently, the investigator has tried to discover the different modelling methods for FFS detection. The researcher has selected a 92 non-FFS and 86 FFS of manufacturing corporations to accomplish the test. From the Bombay stock exchange, the research data were obtained for the dimension of 2008-2011. For the identification of non-FFS and FFS companies, the auditor's report was deliberated. The T-test utilises 31 significant financial proportions. The training dataset is employed to train the model and the trained model is used for classification with better accuracy.

https://doi.org/10.17993/3cemp.2021.100448.77-105

3. METHODOLOGY

The group of financial statements is considered as the input in the text mining systems. Here, the fraud and non-fraud types of financial reports are gathered to classify fake financial reports.

The financial statement fraud discovery includes four steps such as text pre-processing, feature extraction, feature selection, and text classification. the workflow of the proposed approach is shown in Figure 1.

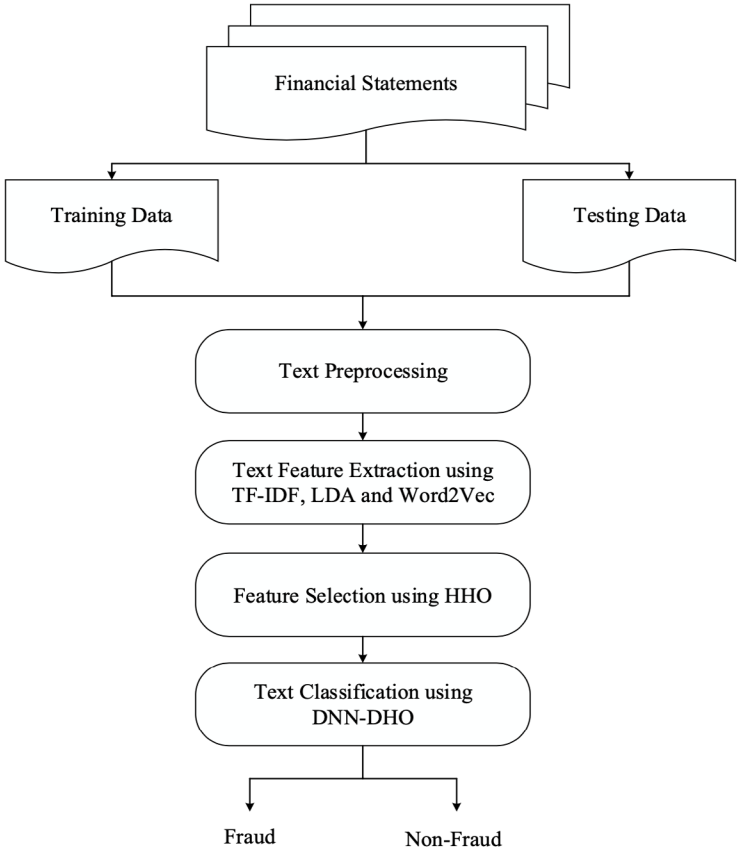

Figure 1. Overall proposed Methodology.
Source: own elaboration.

https://doi.org/10.17993/3cemp.2021.100448.77-105

In-text mining, pre-processing plays a major role. The high quality of the pre-processing step provides better results. The pre-processing step includes the number of roles such as filtering, tokenization, and lemmatization. The words in all documents are transformed into the lower case during pre-processing. Then the TF-IDF, LDA and Word2vec approach is utilized for feature extraction. It describes the text to have a set of measurable dimensions like frequency of words. The process of feature selection is utilized to enhance the performance of a text classifier and also decrease the dimension of the feature. Here, the HHO algorithm is used for feature selection. Finally, the new hybrid classifier of DNN-DHO is proposed to identify the fraud and non-fraud financial statements for classification. In the DNN, weights are updated using the DHO algorithm. This hybrid classifier concept minimizes the error during classification.

3.1. PROBLEM STATEMENT

FSF is the main problem for society. The detection of FSF is a challenging process. FSF is not a victimless corruption, but instead leaves behind actual genuine economic losses that contain workers, shareholders, and investors. A trust in controllers, reduction in self-assurance and reduction in the reliability of financial markets are extensive costs to society. It leads to high transaction costs and minimum efficiency. In developing markets, the challenges of business-related with investing to improve the incentives for handling financial statements and also avoid the taxes in the home country. Recently, different cases of FSF have been increased. Every incidence is a dense disappointment to shareholders, and investors and it expenses the public extremely. So, the construction of an efficient scheme to identify FSF is a major concern.

3.2. TEXT PRE-PROCESSING

It is a significant role and a dangerous phase in text mining. To mining motivating, non-trivial, and information from amorphous text data, a pre-processing method is applied in text mining. The basic units of the fonts, words, and sentences are recognized in this phase and it's delivered to all further

https://doi.org/10.17993/3cemp.2021.100448.77-105

processing phases. The steps of pre-processing contain the number of roles, for example, filtering, lemmatization, and tokenization.

3.2.1. TOKENIZATION

A given text is broken into phrases, words, symbols, or other important components are known as tokens. It may be thrown away particular characters like punctuation marks. The main application of this process is to identify the significant keywords.

3.2.2. FILTERING

This process eliminates the particular words in the documents. The elimination of stop words is a common filtering approach. Stop words are repeatedly utilized common words like 'this', 'are', 'and' etc. They are not applicable in document classification. Therefore, they must be eliminated.

3.2.3. LEMMATIZATION

This process to eliminate inflectional terminations and to return the base form of a word, which is named as the lemma. This process refers to the usage of the dictionary and morphological study of words.

3.3. TEXT FEATURE EXTRACTION

Text feature extraction is the procedure of extracting list of words from the textual data for the feature selection in classifier. In-text classification, it plays a major role because it directly impacts the classification accuracy. The following methods are utilized for extracting features from text data.

3.3.1. TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY (TF-IDF)

TF-IDF is an important weighting method in text mining (Kalra *et al.*, 2019). A word is frequent in the number of times in a text which denoted as the word frequency. To compute the reverse likelihood of finding a word, the IDF approach is employed in a text. The significance of a term in a text is denoted

https://doi.org/10.17993/3cemp.2021.100448.77-105

as TF-IDF within a corpus. Here, a document refers to a financial report, a term refers to a solitary word in a statement, and a corpus refers to the assortment of reports. In a document d, the weight of TF-IDF for a term t is computed by:

$$TF\ (t,d) = \frac{No.\ of\ times\ t\ appears\ in\ d}{Total\ number\ terms\ in\ d} \quad (1)$$

$$IDF\ (t) = \log\left(\frac{Total\ number\ of\ documents}{No.\ of\ documents\ with\ t}\right) \quad (2)$$

$$TF - IDF\ (t,d) = TF\ (t,d) * IDF\ (t) \quad (3)$$

3.3.2. LATENT DIRICHLET ALLOCATION (LDA)

LDA is the topic modelling scheme (Jelodar *et al.*, 2019). It adopts that every text can be defined as a probabilistic distribution over hidden topic. In all documents, the common Dirichlet prior is shared by the topic distribution. A common Dirichlet prior shared by the word distributions of topics. Assumed a corpus D that contains M documents. Each document d having N_d words $d \in 1,\ldots, M$. This method based on the subsequent reproductive procedure:

- From a Dirichlet dissemination with factor β, choose a multinomial spreading φ for a topic $t(t \in 1,\ldots,T)$.
- For document $d(d \in 1,\ldots,M)$, select a multinomial spreading θ_d from a Dirichlet dissemination with factor a.
- Pick a topic z_n from θ_d and take a word w_n from φ_{zn} for a word w_n $(n \in 1,\ldots, N_d)$ in a document d.

Here, the words are only detected variables in documents whereas others are hyper factors (a and β) and hidden variables (φ and θ). The likelihood of perceived data D is calculated by:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)\,p(w_{dn}|z_{dn},\beta)\right)d\theta_d \quad (4)$$

https://doi.org/10.17993/3cemp.2021.100448.77-105

a the spreading of words over topics and β constraints of topic Dirichlet prior are obtained from Dirichlet dissemination. Here the number of topics is defined by T, the number of documents is denoted by M, and the size of the vocabulary is denoted by N. The Dirichlet-multinomial pair is considered as (a,θ) and (β,φ) for the corpus-level topic distributions and the topic-word distributions. The document-level variables are denoted by θ_d, and the word-level variables are represented by w_{dn}.

3.3.3. WORD2VEC

In this process, the depiction of a word as a vector plays a significant role. This process more helpful for discovering antonyms, synonyms, and sentence equivalent with comparable meaning. This process converting the word into a vector form (Wang, Ma, & Zhang, 2016). It contains two different models for constraint updation. One is Continuous Bag of Words (CBOW) and skip-gram. CBOW is used to forecast words utilizing contexts of its environments. The Skip-gram uses a word's data in forecasting of adjacent words. Three layers are used such as input, projection and output are used in both the methods. Here, the CBOW approach is considered as an instance to clarify the working of word2vec.

A sentence S is assumed as:

$S = \{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\} \in R^m$, where w_t refers to the target term. Then the input layer is defined as follows:

$$c(v(w_t)) = \{v(w_{t-2}), v(w_{t-1}), v(w_{t+1}), v(w_{t+2})\} \in R^m \quad (5)$$

Where $c(v(w_t))$ refers to the context of the term $v(w_t)$. Next, the projecting layer is used to construct a contextual vector $v(w_t)$ as follows:

$$v(w_t) = \sum_{i=t-2}^{t+2} c(v(w_i)) \quad (6)$$

A word is considered as a Leaf Node (LN) in a Huffman tree based on its event in the corpus in the output layer. Every word has a single path between the Root Node (RN) and the LN. Using the logistic

https://doi.org/10.17993/3cemp.2021.100448.77-105

model, the likelihood of choosing left or right child can be computed at every node excluding the leaf node which is given by:

$$\begin{aligned} &left\ child: \sigma(v(x_w)^T \theta) = \frac{1}{1+e^{-v(x_w)^T \theta}} \\ &right\ child: 1-\sigma(v(x_w)^T \theta) \end{aligned} \tag{7}$$

At every node, using an invention of likelihoods $p(v(x_w)|c(v(x_w)))$ can be learned in the tree which is given by:

$$p\left(v\left(x_w\right)|c\left(v\left(x_w\right)\right)\right) = \left[\sigma\left(v\left(x_w\right)^T \theta_{j-1}^w\right)\right]^{1-d_j^w} \left[1-\sigma\left(v\left(x_w\right)^T \theta_{j-1}^w\right)^{d_j^w}\right] \tag{8}$$

Here, the j^{th} digit in word $w's$ Huffman code is defined by $d_j^w \in [0,1]$ and any node on the path is denoted as j excluding as the LN.

By enhancing the log-likelihood, the objective purpose can be erudite by (9). Then the gradient descent approach is utilized to improve θ, $v(x_w)$ and its relative words.

$$F = \sum_{w \in c} \log \prod_{j=2}^{n} \left\{\left[\sigma\left(v\left(x_w\right)^T \theta_{j-1}^w\right)\right]^{1-d_j^w} \left[1-\sigma\left(v\left(x_w\right)^T \theta_{j-1}^w\right)^{d_j^w}\right]\right\} \tag{9}$$

3.4. FEATURE SELECTION USING HHO

It is a crucial step for text classification and it is the procedure of choosing a certain subcategory of terms of the training set and these are utilizing for further classification procedure. It also lessens the size of information, improves the classification accuracy by removing noisy features, eliminates overfitting problem and it makes the training faster. The HHO algorithm is introduced in the feature selection process to choose the optimal finest features for text classification. This algorithm analyses the number of features to obtain more relevant features.

3.4.1. HARRIS HAWKS OPTIMIZATION ALGORITHM

HHO is inspired by the behaviour of Harris hawks to discover the prey, surprise pounce, and dissimilar violence methods in the environment (Heidari *et al.*, 2019). The hawks are denoted as the applicant solutions and the finest solution is termed as prey. Using their powerful eyes, the Harris hawks effort to

https://doi.org/10.17993/3cemp.2021.100448.77-105

trail the prey and execute the surprise pounce to hook the prey detected. In this process, three features such as TF-IDF, LDA, and word2vec are taken as input. These three features are not similar to each text. Therefore, the HHO is utilized to select the optimal feature for the classification of text.

Generally, HHO includes the exploration and exploitation stages. The HHO algorithm can be transferred from exploration to exploitation. The exploration behaviour is improved based on the escaping energy of prey (E) and it is given by:

$$E = 2E_0\left(1-\frac{t}{T}\right) \tag{10}$$

$$E_0 = 2r-1 \tag{11}$$

Here the present iteration is denoted by t, the maximum number of iterations is represented by T, the initial energy is defined by E_0 that lies between [-1, 1] and r denoted as a random number in [0, 1].

3.4.1.1. EXPLORATION PHASE

Through the arbitrary position, the location of the hawk is modernized which can be given as:

$$X(t+1) = \begin{cases} X_k(t) - r_1\left|X_k(t) - 2r_2X(t)\right| & q \geq 0.5 \\ \left(X_r(t) - X_m(t)\right) - r_3\left(lb + r_4(ub-lb)\right) & q < 0.5 \end{cases} \tag{12}$$

Here, the location of the hawk is defined by X, the location of the arbitrarily chosen hawk is denoted as X_k, and the location of the prey is defined as X_r. The lower and upper limits of hunt space are signified by lb and ub individually. In the range of [0, 1], the five independent arbitrary numbers are defined by r_1, r_2, r_3, r_4, and q. The ordinary location of the present populace of hawks is defined by X_m and it is given by:

$$X_m(t) = \frac{1}{N}\sum_{n=1}^{N} X_n(t) \tag{13}$$

Here, the n^{th} hawk is denoted as X_n and the number of hawks is defined by N.

https://doi.org/10.17993/3cemp.2021.100448.77-105

3.4.1.2. FITNESS FUNCTION

The fitness value is computed for each hawk and stored for future reference. The fitness function of this feature selection process is computed by:

$$F(x) = Max\{TF - IDF, Word2Vec, LDA\} \tag{14}$$

3.4.1.3. EXPLOTATION PHASE

According to the four dissimilar conditions, the location of the hawk is improved in this process. This process is accomplished only depends on the chance of prey is effectively escaping ($r < 0.5$) or not effectively escaping ($r \geq 0.5$) beforehand surprise bounce and the escaping energy of prey (E).

- Soft Besiege

If $|E| \geq 0.5$ and $r \geq 0.5$, this stage only occurs. Here, the location of the hawk is updated by the subsequent expression:

$$X(t+1) = \Delta X(t) - E\left[JX_r(t) - X(t)\right] \tag{15}$$

Here, the dissimilarity between the current hawk and the position of the prey is denoted as ΔX and the jump strength is denoted by J. Both parameters can be defined as:

$$\Delta X(t) = X_r(t) - X(t) \tag{16}$$

$$J = 2\left(1 - r_5\right) \tag{17}$$

Where r_5 is a constant value in the range of 0 and 1 that changes unevenly in every single iteration.

- Hard Besiege

If $|E| < 0.5$ and $r \geq 0.5$, this phase only happens. Here, the location of the hawk is updated by the following expression:

https://doi.org/10.17993/3cemp.2021.100448.77-105

$$X(t+1) = X_r(t) - E\left|\Delta X(t)\right| \tag{18}$$

- Soft Besiege with Progressive Rapid Dives

If $|E| \geq 0.5$ and $r < 0.5$, this stage is happened. The hawk gradually picks the finest probable dive to catch the prey. Here, the two different solutions are produced by,

$$Y = X_r(t) - E\left[JX_r(t) - X(t)\right] \tag{19}$$

$$Z = Y + \alpha \times Levy(D) \tag{20}$$

Here, the newly produced hawks are denoted by Y and Z. the total number of dimensions is denoted as D, a is an arbitrary vector and Levy is the function of levy flight which is given by:

$$Levy(x) = 0.01 \times \frac{\mu\sigma}{|v|^{1/\beta}} \tag{21}$$

Here, u and v are the self-governing arbitrary numbers produced from the standard distribution and σ is given by:

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}} \right)^{\frac{1}{\beta}} \tag{22}$$

Where β is a constant value fixed to 1.5. Here, the location of the hawk is reorganized by:

$$X(t+1) = \begin{cases} Y & if \ F(Y) < F(X(t)) \\ Z & if \ F(Z) < F(X(t)) \end{cases} \tag{23}$$

Where the fitness function is defined as $F(.)$, Y and Z are two different solutions gained from Equations (19) and (20).

https://doi.org/10.17993/3cemp.2021.100448.77-105

- Hard Besiege with Progressive Rapid Dives

If $|E| < 0.5$ and $r < 0.5$, this process is occurred. The two different solutions are made by:

$$Y = X_r(t) - E\left[JX_r(t) - X_m(t)\right] \tag{24}$$

$$Z = Y + \alpha \times Levy(D) \tag{25}$$

The location of the hawk is updated by:

$$X(t+1) = \begin{cases} Y & if\ F(Y) < F(X(t)) \\ Z & if\ F(Z) < F(X(t)) \end{cases} \tag{26}$$

Where Y and Z are two fresh solutions achieved from Equations (24) and (25).

3.5. OPTIMIZED DNN BASED CLASSIFICATION USING DHO

The structure of DNN includes the input layer, hidden layers, and output layer as exposed in Figure 2. By the exertion of weight fitness, the network is constructed. DNN updates the weight value in the hidden layer using the DHO algorithm (Brammya *et al.*, 2019). Owing to the improved training repetitions, this system frequently fits the considered training information's judgment border. The total quantity of nodes is evaluated in the hidden layers which are given as:

$$n = \sqrt{a+b} + c \tag{27}$$

https://doi.org/10.17993/3cemp.2021.100448.77-105

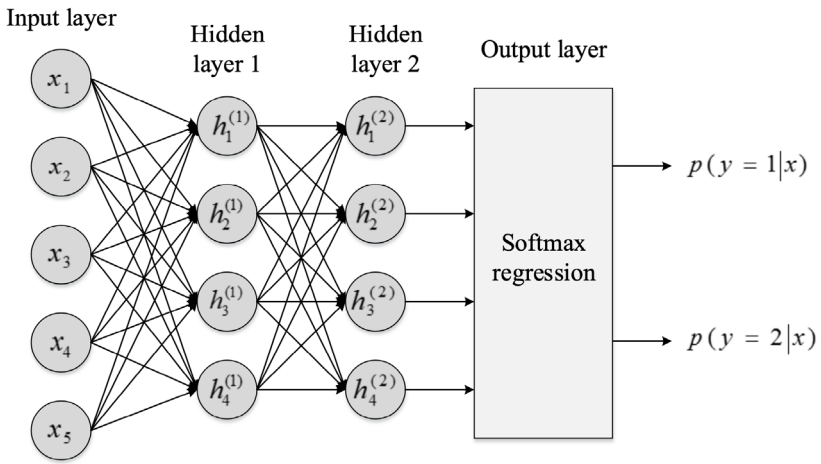

Figure 2. DNN with SoftMax regression.
Source: own elaboration.

Here, the sum of the hidden layer is n, the layer of input is a, the layer of output is defined as b and c is a constant where $0 \leq c \geq 1$. The sigmoid utility is used as an activation function for empowering the non-linear capability which is computed as:

$$S = \frac{1}{1+e^{-x}} \tag{28}$$

The input information of the system is considered as and the mapping function is defined as M_f.

$$M_f = sigm(\omega_i x + \beta_i) \tag{29}$$

In this w is weight matrix, and β is bias between output and hidden layer. A data model (x, l) can be taken and the loss form computed as:

$$S(W_s, b_s; x, l) = \frac{1}{2m}\sum_{j=1}^{m}\left\| h_j(W_s, b_s; x) - l_j \right\|_2^2 \tag{30}$$

https://doi.org/10.17993/3cemp.2021.100448.77-105

Here, W_s and b_s are bias subsets, hidden layer nodes are m the sum of neurons in the hidden layer is signified as m. The Cross-Entropy (CE) for the testing and training of the model is taken as loss form for the deep neural network. This can be estimated as:

$$C_E = \frac{1}{n}\sum_{k=1}^{n}\left[Y_k \log \hat{Y}_k + (1-Y_k)\log(1-\hat{Y}_k)\right] \quad (31)$$

Here sample of training is n, the k^{th} output is y_k from training set and the expected k^{th} output is $\hat{Y}_k$. The network weight value is estimated by the DHO method.

Then the old and fresh solutions are compared. Only the best solutions are considered for the next iteration. Furthermore, it simply needs the alteration of the population dimensions. The number of iterations updates the calculation in only one stage.

In DHO the two hunters one is leader and other is successor must be at their best position. For this they update their angle and position hunt the deer. the leader updates his angle and position as:

$$Y_{i+1} = Y^{lead} - p\left|\cos(v) \times Y^{lead} - Y_i\right| \quad (32)$$

Where, Y_i is the current position, Y_{i+1} is the next position, p is a random number belongs to [0, 2]. Leader present position is Y^{lead} from a present population.

The position can be updated by successor position as:

$$Y_{i+1} = Y^{successor} - X.p.\left|L \times Y^{successor} - Y_i\right| \quad (33)$$

Where successor position is $Y^{successor}$. The coefficient factors can be calculated as:

$$X = \tfrac{1}{4}\log\left(i + \frac{1}{i_{\max}}\right)b \quad (34)$$

$$L=2.c \quad (35)$$

https://doi.org/10.17993/3cemp.2021.100448.77-105

i_{max} is maximum repetition, b is an arbitrary number between -1 to 1, here c is a number from 0 to 1. The mean value of leader and successor can be used for weight updation:

$$Mean = \frac{Position\ of\ the\ leader + Position\ of\ the\ successor}{2} \tag{36}$$

The current and earlier solution are compared. It will replace the earlier solution if the earlier solution is enhanced otherwise, it will keep the earlier solution. This procedure is frequent up to the end condition is satisfied. DHO algorithm for weight estimation is shown in Figure 3.

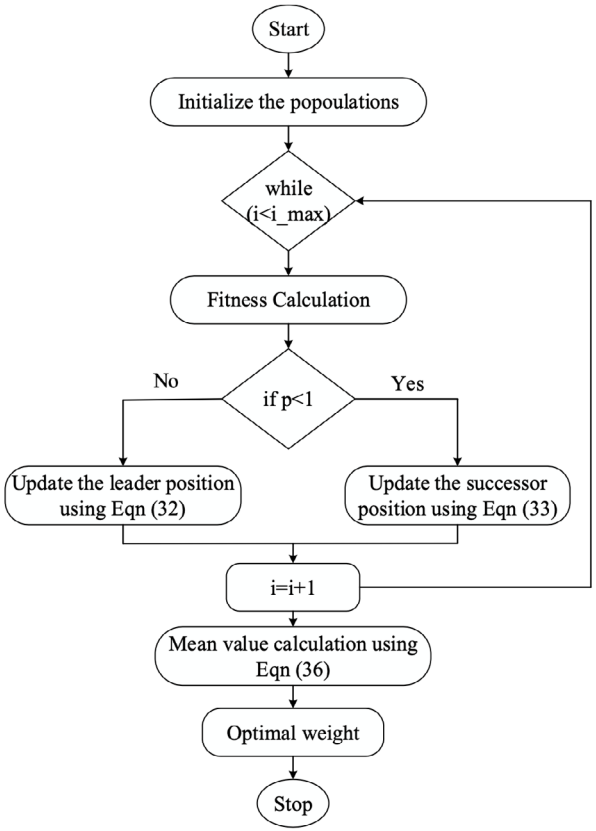

Figure 3. DHO for weight updation.
Source: own elaboration.

https://doi.org/10.17993/3cemp.2021.100448.77-105

3.6. DATASET DESCRIPTION

The standard datasets are utilized for FSF detection. The full dataset is taken from the link https://surfdrive.surf.nl/files/index.php/s/m34LCElefSj6M8y. Here annual reports are stored in two zip files. One file contains the annual reports in the 'fraud' category and other files in the 'no fraud' category. 1646 statements are included in the datasets. It contains 1319 no fraud statements and 327 fraud statements. For this work 70% data is used for training purpose and 30% data is used for testing purpose.

4. RESULTS

DNN-DHO methods are proposed here to optimize the DNN model for detection of FSF. Different classifiers such as DNN, K-nearest neighbour (KNN) SVM, backpropagation neural network (BP-NN), classification and regression tree (CART), Bayes classifier (Bayes), and logistic regression (LR) are compared with the proposed approach and a comparative evaluation performance is made.

4.1 PERFORMANCE ANALYSIS

The proposed (DNN-DHO) approach is executed on the financial statement dataset. The proposed scheme correctly identifies the fraud and non-fraud statement. Initially, the pre-processing step is executed on the text data. It includes the number of tasks such as tokenization, filtering, and lemmatization. These stages are performed on the text. After that, different feature extraction methods like TF-IDF, LDA, and word2vec are utilized. Then, the HHO algorithm selects the optimal. The features selected optimally is used by DNN-DHO algorithm to classify the financial statement. The proposed method provides better outcomes compared to the standard classifiers like CART, DNN, SVM, NB, LR, BP-NN, and KNN. The performance metrics are evaluated for different classifiers. The accuracy obtained by this method is better (96%) than other standard classifiers.

https://doi.org/10.17993/3cemp.2021.100448.77-105

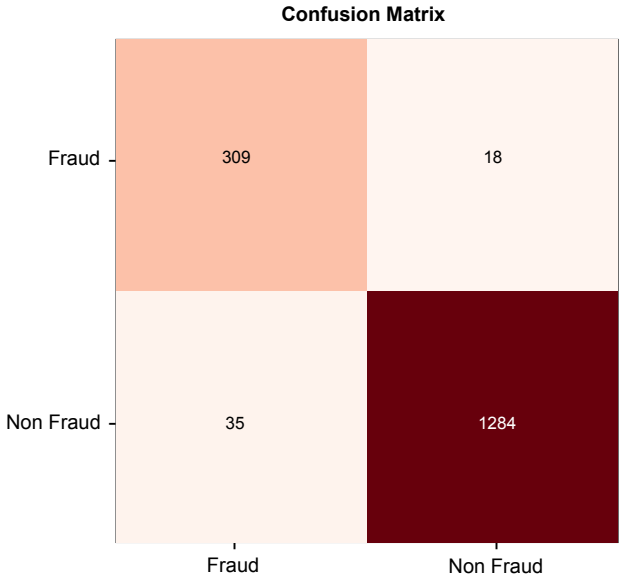

Figure 4. Confusion Matrix.
Source: own elaboration.

The confusion matrix obtained by the proposed method is shown in Figure 4. The financial statements consist of 327 fraud statements and 1319 non-fraud statements. In 327 fraud statements, 309 financial statements are correctly identified as a fraud statement remaining 18 financial statements are wrongly identified as a non-fraud statement. Similarly, in the 1319 non-fraud statements, 1284 financial statements are correctly identified as a non-fraud statement remaining 35 financial statements are wrongly identified as a fraud statement. Therefore, the proposed scheme properly identifies the fraud or non-fraud in the financial statements. The accuracy evaluation is shown in Figure 5.

https://doi.org/10.17993/3cemp.2021.100448.77-105

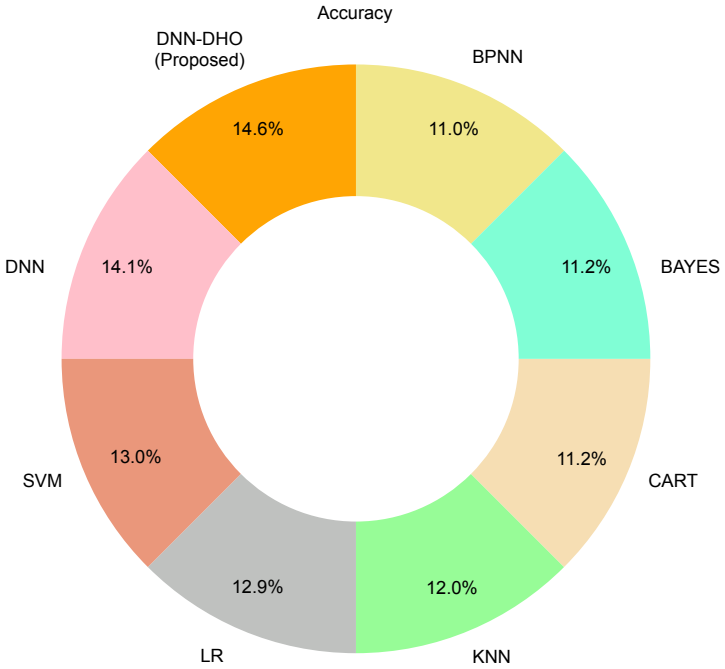

Figure 5. Accuracy evaluation.
Source: own elaboration.

The outcomes of the proposed classifier and standard classifiers like SVM, CART, Bayes, BP-NN, DNN, LR, and KNN is shown in Table 1 and Figure 6. The performance of accuracy is 96% for the DNN-DHO, 93% for DNN, 86% for SVM, 74% for CART, 73% for BP-NN, 85% for LR, 74% for Naïve Bayes and 79% for KNN. DNN-DHO method outperforms in all other parameters also. The proposed approach is better in comparison to the existing classifiers for classification of FSF.

Table 1. performance comparison among DNN-DHO classifier and others.

Performance parameters	DNN-DHO (Proposed)	DNN	SVM	LR	KNN	CART	Bayes	BPNN
Accuracy	0.9678	0.93	0.86	0.85	0.79	0.74	0.74	0.73
Sensitivity	0.9734	0.94	0.89	0.87	0.82	0.83	0.82	0.81
FPR	0.055	0.065	0.16	0.16	0.23	0.34	0.34	0.35

https://doi.org/10.17993/3cemp.2021.100448.77-105

FNR	0.0265	0.06	0.106	0.12	0.17	0.16	0.17	0.18
Precision	0.9861	0.93	0.84	0.84	0.77	0.71	0.70	0.69
F1 score	0.9797	0.93	0.87	0.85	0.80	0.76	0.76	0.75
Specificity	0.9449	0.935	0.84	0.83	0.76	0.66	0.65	0.64
BER	0.0321	0.0625	0.133	0.14	0.20	0.25	0.25	0.26
AUC	0.9423	0.9178	0.8725	0.86	0.79	0.78	0.75	0.72

Source: own elaboration.

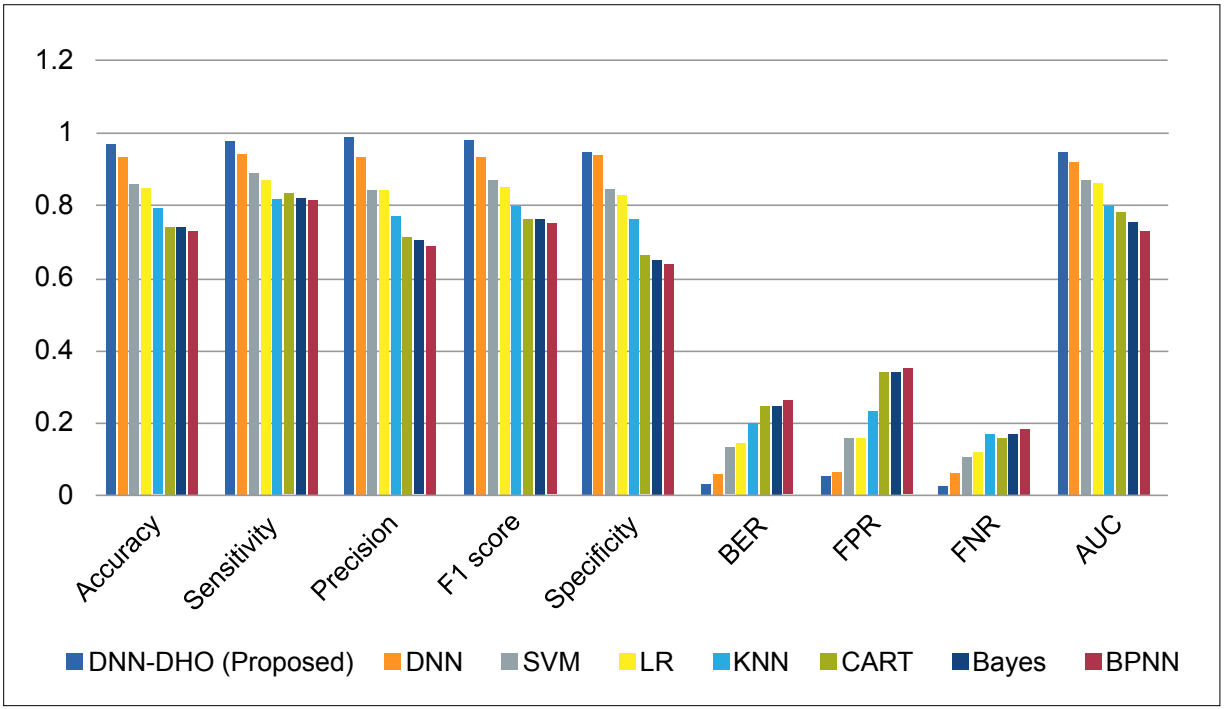

Figure 6. performance comparison among DNN-DHO classifier and others.
Source: own elaboration.

5. CONCLUSIONS

In this paper, an optimized deep neural network based FSF discovery in text mining has been proposed. The model of fraud detection initiates with an assortment of financial reports for both fraud and no-

https://doi.org/10.17993/3cemp.2021.100448.77-105

fraud administrations. The pre-processing stage is performed through lemmatization, filtering, and tokenization. Then the TF-IDF, LDA and word2vec approach is used for mining the data concealed in the document for fraud and no-fraud administrations. Further, the HHO procedure is utilized to select the finest features. Then the DNN-DHO classifier utilises these features with a SoftMax classifier for classification of fraud and no-fraud statements. In the classification process, the weight of the whole network is updated by the DHO algorithm. The outcomes shows that the proposed method is the best model for detecting FSF. The accuracy (96%), Sensitivity (97%), precision (98%), F1 score (97%), Specificity (94%), BER (0.03), FPR (0.05), FNR (0.026) and AUC is 0.94 are calculated for the developed method and it's compared to the existing classifiers. The proposed approach is to provide the best performance results than other classifiers of BP-NN, DNN, CART, SVM, LR, KNN, and Bayes.

REFERENCES

Asare, S. K., Wright, A., & Zimbelman, M. F. (2015). Challenges facing auditors in detecting financial statement fraud: Insights from fraud investigations. *Journal of Forensic and Investigative Accounting, 7*(2), 63-111. http://web.nacva.com/JFIA/Issues/JFIA-2015-2_4.pdf

Brammya, G., Praveena, S., Ninu, N. S., Ramya, R., Rajakumar, B. R., & Binu, D. (2019). Deer Hunting Optimization Algorithm: A New Nature-Inspired Meta-heuristic Paradigm. *The Computer Journal, bxy133*. https://doi.org/10.1093/comjnl/bxy133

Chen, Y.-J., Liou, W.-C., Chen, Y.-M., & Wu, J.-H. (2019). Fraud detection for financial statements of business groups. *International Journal of Accounting Information Systems, 32*(C), 1-23. https://ideas.repec.org/a/eee/ijoais/v32y2019icp1-23.html

Chi, D.-J., Chu, C.-C., & Chen, D. (2019). Applying Support Vector Machine, C5. 0, and CHAID to the Detection of Financial Statements Frauds. In *International Conference on Intelligent Computing*, pp. 327-336. Springer, Cham.

Dalnial, H., Kamaluddin, A., Sanusi, Z. M., & Khairuddin, K. S. (2014). Detecting fraudulent financial reporting through financial statement analysis. *Journal of Advanced Management Science, 2*(1), 17-22. http://www.joams.com/index.php?m=content&c=index&a=show&catid=36&id=108

Dong, W., Liao, S., & Liang, L. (2016). Financial Statement Fraud Detection using Text Mining: A Systemic Functional Linguistics Theory Perspective. In *Pacific Asia Conference On Information Systems (PACIS)*, p. 188. https://core.ac.uk/download/pdf/301369656.pdf

Gray, G. L., & Debreceny, S. R. (2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. *International Journal of Accounting Information Systems, 15*(4), 357-380. https://doi.org/10.1016/j.accinf.2014.05.006

Gupta, R., & Gill, N. S. (2012a). A data mining framework for prevention and detection of financial statement fraud. *International Journal of Computer Applications, 50*(8). https://research.ijcaonline.org/volume50/number8/pxc3880889.pdf

Gupta, R., & Gill, N. S. (2012b). Financial statement fraud detection using text mining. *International Journal of Advanced Computer Science and Applications (IJACSA), 3*(12). http://dx.doi.org/10.14569/IJACSA.2012.031230

Gupta, R., & Gill, N. S. (2012c). Prevention and detection of financial statement fraud–An implementation of data mining framework. *International Journal of Advanced Computer Science and Applications (IJACSA), 3*(8). http://dx.doi.org/10.14569/IJACSA.2012.030825

Hajek, P. (2019). Interpretable Fuzzy Rule-Based Systems for Detecting Financial Statement Fraud. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 425-436. Springer, Cham.

Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud–A comparative study of machine learning methods. *Knowledge-Based Systems, 128,* 139-152. https://doi.org/10.1016/j.knosys.2017.05.001

Heidari, A. A., Mirjalili, S., faris, H., Aljarah, I., Mafarja, M., & Chen, H. (2019). Harris hawks optimization: Algorithm and applications. *Future Generation Computer Systems 97,* 849-872. https://doi.org/10.1016/j.future.2019.02.028

Jan, C.-L. (2018). An effective financial statement fraud detection model for the sustainable development of financial markets: Evidence from Taiwan. *Sustainability, 10*(2), 513. https://doi.org/10.3390/su10020513

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169-15211. https://arxiv.org/abs/1711.04305

Kalra, S., Li, L., & Tizhoosh, H. R. (2019). Automatic Classification of Pathology Reports using TF-IDF Features. *arXiv preprint arXiv:1903.07406.* https://arxiv.org/abs/1903.07406

Kanapickienė, R., & Grundienė, Ž. (2015). The model of fraud detection in financial statements by means of financial ratios. *Procedia-Social and Behavioral Sciences, 213,* 321-327. https://doi.org/10.1016/j.sbspro.2015.11.545

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in the financial domain. Knowledge-Based Systems, 114, 128-147. https://doi.org/10.1016/j.knosys.2016.10.003

Lin, C., Chiu, A., Huang, S.Y., & Yen, D.C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems, 89,* 459-470. https://www.semanticscholar.org/paper/Detecting-the-financial-statement-fraud%3A-The-of-the-Lin-Chiu/48bc08514070341439e382f887faba42b21212d9

Manurung, D. T. H., & Hardika, A. L. (2015). Analysis of factors that influence financial statement fraud in the perspective fraud diamond: Empirical study on banking companies listed on the Indonesia Stock Exchange year 2012 to 2014. In *International Conference on Accounting Studies (ICAS)*, 279-286. https://core.ac.uk/download/pdf/42984276.pdf

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of the literature. *Decision support systems, 50*(3), 559-569. https://doi.org/10.1016/j.dss.2010.08.006

Omar, N. B., Koya, R. K., Sanusi, Z. M., & Shafie, N. A. (2014). Financial Statement Fraud: A Case Examination Using Beneish Model and Ratio Analysis. *International journal trade, economics and finance, 5,* 184-186. https://www.semanticscholar.org/paper/Financial-Statement-Fraud%3A-A-Case-Examination-Using-Omar-Koya/75657feb5f290f2c5447eb71573b3b6753c17bfb

Patel, H., Parikh, S., Patel, A., & Parikh, A. (2019). An application of ensemble random forest classifier for detecting financial statement manipulation of Indian listed companies. In *Advances in Intelligent Systems and Computing, Recent Developments in Machine Learning and Data Analytics.* Springer Proceedings. https://www.researchgate.net/profile/Satyen-Parikh/publication/327604170_An_Application_of_Ensemble_Random_Forest_Classifier_for_Detecting_Financial_Statement_Manipulation_of_Indian_Listed_Companies_IC3_2018/links/5e8b631e299bf1307983c98e/An-Application-of-Ensemble-Random-Forest-Classifier-for-Detecting-Financial-Statement-Manipulation-of-Indian-Listed-Companies-IC3-2018.pdf

Perols, J. L., & Lougee, B. A. (2011). The relation between earnings management and financial statement fraud. *Advances in Accounting, 27*(1), 39-53. https://doi.org/10.1016/j.adiac.2010.10.004

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50(*2), 491-500. https://doi.org/10.1016/j.dss.2010.11.006

Rezaee, Z., & Kedia, B. L. (2012). Role of corporate governance participants in preventing and participants in preventing and detecting financial statement fraud. *Journal of Forensic & Investigative Accounting, 4*(2).

Song, X.-P., Hu, Z.-H., Du, J.-G., & Sheng, Z.-H. (2014). Application of machine learning methods to risk assessment of financial statement fraud: evidence from China. *Journal of Forecasting, 33*(8), 611-626. https://doi.org/10.1002/for.2294

Temponeras, G. S., Alexandropoulos, S. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Financial Fraudulent Statements Detection through a Deep Dense Artificial Neural Network. In *10th International Conference on Information, Intelligence, Systems, and Applications (IISA)*, pp. 1-5. IEEE. https://ieeexplore.ieee.org/abstract/document/8900741

Throckmorton, C. S., Mayew, W. J., Venkatachalam, M., & Collins, L. M. (2015). Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems, 74,* 78-87. https://doi.org/10.1016/j.dss.2015.04.006

Wang, Z., Ma, L., & Zhang, Y. (2016). A Hybrid Document Feature Extraction Method Using Latent Dirichlet Allocation and Word2Vec. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, 98-103. https://www.semanticscholar.org/paper/A-Hybrid-Document-Feature-Extraction-Method-Using-Wang-Ma/840894b784378fe64ef977c44db759b8aa0527cf

https://doi.org/10.17993/3cemp.2021.100448.77-105