IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 28, 2022, accepted February 9, 2022, date of publication February 22, 2022, date of current version March 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3153478

An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning

WU XIUGUO AND DU SHENGYONG

School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan, Shandong 250014, China

Corresponding author: Wu Xiuguo (xiuguow@sdufe.edu.cn)

This work was supported in part by the National Social Science Foundation of China under Grant 20BJY033, and in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2020MG031.

ABSTRACT Financial fraud has extremely damaged the sustainable growth of financial markets as a serious problem worldwide. Nevertheless, it is fairly challenging to identify frauds with highly imbalanced dataset because ratio of non-fraud companies is very high compared to fraudulent ones. Intelligent financial statement fraud detection systems have therefore been developed to support decision-making for the stakeholders. However, most of current approaches only considered the quantitative part of the financial statement ratios while there has been less usage of the textual information for classifying, especially those related comments in Chinese. As such, this paper aims to develop an enhanced system for detecting financial fraud using a state-of-the-art deep learning models based on combination of numerical features that derived from financial statement and textual data in managerial comments of 5130 Chinese listed companies' annual reports. First, we construct financial index system including both financial and non-financial indices that previous researches usually excluded. Then the textual features in MD&A section of Chinese listed company's annual reports are extracted using word vector. After that, powerful deep learning models are employed and their performances are compared with numeric data, textual data and combination of them, respectively. The empirical results show great performance improvement of the proposed deep learning methods against traditional machine learning methods, and LSTM, GRU approaches work with testing samples in correct classification rates of 94.98% and 94.62%, indicating that the extracted textual features of MD&A section exhibit promising classification results and substantially reinforce financial fraud detection.

INDEX TERMS Fraud detection, feature selection, deep learning, text analytics, LSTM.

I. INTRODUCTION

With the boom of the securities market in last decades, more and more companies raise capital and expand the operation scale through listing, especially in fast growing counties like China. Accompanied by financial market development, fraudulent financial reports have cast rapidly, and have caused dramatic losses to shareholders with negative impacts on capital markets [1], [2]. The Enron scandal in the U.S. in 2001 and the global financial crisis spanning 2008–2009 have severely damaged the world economy [3]. In China, the number of criminals involved with fraudulent activities in 2019 is more than 961 with a value of more than $8 billion [4]. Although there are minor variations in its definition, a financial statement fraud is referred as "deliberate fraud committed by management that injures investors and creditors through misleading financial statements" [2]. Generally speaking, the main reason for fraud is due to the inaccurate reports of CPAs and auditors. In addition, companies with rapid growth may exceed the monitoring process ability to provide appropriate supervision. According to report issued in 2020, only a limited number of fraud cases were identified by internal and external auditors with rates of 14% and 5%, respectively [5]. As a result, effective detecting financial fraud has always been an important but rather challenging task for accounting and auditing professionals given that the economic and social consequences can be massive [5], [6]. However, traditional manual detection

The associate editor coordinating the review of this manuscript and approving it for publication was Rosalia Maglietta.

approaches are not only tedious, inaccurate and complex, but also impractical for the management of big semi-structured and unstructured financial data these days. In fact, an experienced analyst draws his/her conclusions depending on not only the numerical data from listed company's financial statement, but also any other information related to the company, such as textual analysis in managerial comments. To our knowledge, only a few researchers have utilized text data in financial statements fraud detection.

According to Securities Law and Measures for the Administration of Listed Companies Information Disclosure of CSRC (China Securities Regulatory Commission), all the listed companies must release their annual reports and audit opinions before April 30 annually, explaining their loans, profits, expenses and incomes. Management's Discussion and Analysis (MD&A) is an important part in section IV of Chinese listed companies' annual reports, called Business Situation Discussion & Analysis (BSD&A), which is usually regard as the textual explanations for those numerical data. Some researchers have verified the prediction ability of MD&A section in financial statements for detecting financial fraud [7], [8]. Nevertheless, the financial statement detection using textual content in MD&A still lacks framework with comprehensive textual features specifics, especially for those text data written in Chinese. It is actually the primary motivation of this study, and the first objective is to provide textual classification framework in MD&A and test out the effectiveness of fraud detection tools by Chinese listed companies' annual reports.

Meanwhile, as has shown that the newest technologies can effectively improve the information management efficiency and quality, particularly in the context that the integration of structure and unstructured data is generally common. Deep learning (DL) is a sub-field of machine learning that models high-levels data abstractions through hierarchical learning layer [9]. As a powerful tool for modeling, it can quickly and effectively reveal the facts that hidden in large amounts of data, and has been widely used to solve complex problems in various area. More recently, deep learning has begun to appear in financial research and has gained some achievements to predict companies' financial risk probability. Unfortunately, current researches most focused on the solely digital data using deep learning methods, such as CNN (Convolution Neural Network) [10] and RNN (Recurrent Neural Network) [11]. In addition, the fraud factors selections were merely according to experts' own experience and specialized knowledge. To solve these limitations, the second objective of this paper is to expand the scope of fraud indicators, providing a combination of digital financial data and textual data as input of deep learning models. Additionally, we will discuss whether or not the combination input model adopting those novel techniques agrees with the previous results obtained, using metrics like accuracy, sensitivity and area under the receiver operating curve(AUC).

Based on these analyses above, it is important to improve the financial statement fraud predictive power using as much data as possible in the listed company's annual reports. And this study constructs a state-of-art fraud detection framework by a combination of numeric and textual data as input with powerful deep learning solution in the era of big data. Furthermore, both financial and non-financial variables (also known as corporate governance variables) are used as the input to detect the signs of financial statement fraud. And the main contributions of this paper are as follows:

(1) A novel multi-dimensional financial fraud factors index system derived from financial information and managerial comments in Chinese listed companies' annual reports, is proposed for Chinese listed companies;

(2) A Chinese textual data mining framework for fraud detection from MD&A in listed companies' annual report using state-of-art deep learning models, is presented;

(3) An enhanced system for detecting financial fraud with combination of numerical features that derived from financial information and textual data in managerial comments, is given;

(4) About 5130 annual reports of Chinese listed companies are mined with deep learning methods, and empirical results suggest the better feasibility and effectiveness of proposed approach.

The rest of the paper is organized as follows: firstly, related works carried out by the researchers are discussed followed by research methodologies used in this study. Section IV presents a detailed description of the fraud detection indicators used in the financial statements mining. Section V gives the classification results by means of empirical analysis and compares the performance with a set of machine learning models. After that, Section VI discusses and analyses the textual data mining and imbalance data treatment in our study. Finally, Section VII concludes the paper with a summary and provides directions for future research.

II. LITERATURE REVIEW

As a hot research topic in recent years, most of the previous studies about financial statements fraud detection mainly involved financial fraud indicators selection and financial fraud detecting techniques.

A. FINANCIAL FRAUD DETECTION FACTORS SELECTION

As a company's basic document, financial statement is an important and essential part in annual report, reflecting its financial status in the recent past and the near future. Nevertheless, it is difficult and cumbersome to manually find accounting irregularities and financial fraud information from the financial statement itself at the surface level. In the past years, many researchers have applied various approaches to detect frauds using financial statement, such as analytical procedures, ratio analysis, score propagation over an auction network and checklists to improve the fraud detection quality and efficiency [12]. However, the majority of existed studies usually result in too fraud risk factors and cannot efficiently and accurately identify those frauds. How to identify some key fraud factors that relevant for detection of financial

statement frauds, and rank the importance of those fraud factors have become paramount issues. These factors mainly include Z-score [9], accounts receivables [10], inventories [11], gross margins [13] and so on. Beyond that, many other financial ratios are also used for fraud detection, such as net profits/total assets, working capitals/total assets, net profits/sales, current assets/current liabilities [14], [15]. Reference [15] employed 32 factors as financial fraud attributes, including pressure/incentive dimension, opportunity dimension and attitude/rationalization dimension.

The same study further adopts AHP in calculating the weightings of individual measurement items, with pressure/incentive as the highest weight. Reference [16] introduced four features of delisting company, including debt-equity ratio, accounts receivable turnover ratio, operating profit ratio and retained earnings ratio to total assets. In addition, a low ratio of selling and administrative expenditure to revenues was reported for firms engaged in revenue fraud [17]. From the existing research literature, as can be seen that the results of fraud factors might be different from the real situations. And the selected collection of financial variables should cover as many aspects as possible in order to identify the various type of financial reporting frauds. However, most of them tend to select only part of the financial items, which is not sufficient in identifying the financial fraud. In addition, there is no non-numeric data involved in the analysis, which are related to the corporate governance structure. Although numerical financial variables are very important and essential for the detection of fraud, it is wise to enhance the performance through the inclusion of other types of data, such as managerial comments in annual report. To solve these problems, this study aims to develop an enhanced system for detecting financial fraud based on combination of numerical features that derived from financial statement and textual data in managerial comments of 5130 Chinese listed companies' annual reports.

B. FINANCIAL FRAUD DETECTION TECHNIQUES

Over the years, various methods for financial fraud detection are always accompanied by the development of information technology. Statistical methods have been used to classify and detect frauds, where financial indicators are the core and fundamental part in prediction, as have discussed in the previous sub-section. More recently, data mining techniques are regard as an effective tool to extract and discover the hidden truths behind the very large quantities of data. And some researchers have gone into addressing fraud detection using predictive and classification technologies [18], [19]. These models include logistic regression (LR) [20], [21], support vector machine (SVM) [22], random forest (RF) [23] and artificial neural network (ANN) [24]. Still, most current financial fraud detection researches limited their investigations only to numerical data in financial statements, ignoring the textual data in the listed company's annual report, especially those related comments written in Chinese. In addition, due to deliberate concealment, fraudulent financial data could hardly be distinguished from authentic data in practice using traditional machine learning methods. Meanwhile, machine learning techniques are also used to detect financial frauds and no fraud detection systems have been able to offer great efficiency to date [25]–[37]. Table 1 depicts the status-quo in the field of financial fraud detection along six dimensions: research reference, the technique utilized, the type of data, the country of study, the predictive performance in terms of classification accuracy and other metrics.

As is shown in Table 1, only a few studies tried to resolve financial restatement problem using deep learning techniques while multi-layer perceptron (MLP), decision tree (DT), naive Bayesian (NB) and SVM are widely adopted as classification models. The majority of existing studies fed solely numerical values as input to the algorithm. Unfortunately, due to deliberate concealment and accounting shenanigans, fraudulent financial data could hardly be distinguished from authentic data. In this way, few researchers used textual data for classifying the financial statement fraud, including corporate conference calls, media reports and annual reports [31], [32]. Reference [31] presented a synergy for extracting both word-level features and document-level features by integrating three analysis methods under the guidance of SFL theory, and reached average prediction accuracy at 82.36 percent. Although they have experimented with linguistic variables, the majority of those approaches only examined the relation between linguistic aspects and fraudulent actions. In this way, approaches using textual content still lacks a systematic and theoretical analysis framework to predict fraud.

One recent literature has applied deep learning techniques to the fraud detection task employed a hierarchical attention network (HAN) with a long short-term memory (LSTM) encoder to extract the text features from the MD&A section of annual reports [37]. This study is closest to our work as they also combined financial numerical and textual data as inputs and employed a variety of classification models, as shown in Table 1. Despite these similarities, their textual data mining approaches still differ widely for the main reason that the Chinese text mining is ever more complex compared to English. If we use the same method as in previous research, some important information will be lost for fraud detection and thus decrease algorithms' detection accuracy. Moreover, they were not targeted at evaluating the textual content of listed companies' annual reports. Therefore, it did not include modern NLP approaches such as deep learning-based feature extraction.

FIN-financial data, LING-linguistic data, TXT-text data, BBN-Bayesian belief network, NB-Naive Bayer, DTNB-NB with the induction of decision table, CART-classification and regression tree, LMT-logistic model trees, MLP-multi-layer network, GP-genetic programming, LR-logistic regression, DNN-deep neural network, PNN-probabilistic neural network, RF-random forest, SVM-support vector machine, Acc-Accuracy, AUC-area under the ROC curve, MC-misclassification cost, TPR-true

TABLE 1. Analysis of comparisons with financial statement fraud methods.

Study	Data(fraud/non-fraud)	Country	Features	Classifiers(Acc(%))	Used metrics
P. Ravisankar, et al.(2011)[25]	101/101	China	FIN	PNN(98.1), GP(94.1), GMDH(93.0), DNN(78.4), Exhaustive pruning NN(77.1)	Acc, TPR, TNR
P. Hajek & R. Henriques(2017)[26]	311/311	US	FIN+LING	BBN(90.3), DTNB(89.5), RF(87.5),BAG(87.1), JRIP(87.0),CART(86.2), C4.5(86.1), LMT(85.4),SVM(78.0)	Acc, TPR, MC, F-score, AUC
C. C. Lin, et al.(2015)[27]	129/447	China(Taiwan)	FIN	ANNs(92.8), CART(90.3%),LR(88.5)	Acc, FPR, FNR, F-score
K. K. Tangod & G. H. Kulkarni(2015) [28]	/	/	FIN	K-Means(79.0), Multi-Level Feed Forward network(MLEF)(73.0)	Acc, Specifity, Senstivity
J. Perols(2011)[29]	51/15934	US	FIN	SVM(MC 0.0025), LR(0.0026), C4.5(0.0028),DNN(0.0030)	Fraud probability and MC
F. H. Glancy & S. B. Yadav (2011)[30]	11/20	US	TXT	Hierarchical clustering (83.9)	TP, TN, FP, FN, p-value
W. Dong, et al.(2016)[31]	805/805	US	TXT	SVM+SFL(82.4)	Acc, Precision, F1-Score, FNR
L. Purda & D. Skillicom(2015)[32]	1407/4708	US	TXT	SVM(AUC 89.0)	AUC, Fraud Probability
P. M. Dechow, et al(2011)[33]	293/79358	US	FIN	LR(63.7)	Acc, TPR, FPR， min F-score
J. Yao et al.(2018)[34]	120/120	China	FIN+TXT	SVM(7107), RF(69.2), DT(69.6), ANN(70.8), LR(70.8)	Acc
A. Byungdae & S. Yongmoo (2020)[35]	1591/31628	Korea	FIN	DT (74.8), RF(78.1) Bagging of DTs (78.0), LR(72.0), SVM(71.4), ANN(78.5) Modified RF (MRF)(79.9)	Acc, Precision, F1-Score
C．Chimonaki, et al.(2019)[36]	231/2469	Greece	FIN	KNN(89.0), NB(68.0)	Acc, Sensitivity, Precision
P. Craja, et al. (2020)[37]	208/7341	US	FIN+TXT	HAN(84.6), ANN(90.5), SVM(82.8), XGB(90.8), RF(87.4), GPT-2+Attn(69.3)	Acc, AUC, F1-Score1, Sensitivity
This work	244/5130	China	FIN+TXT	RNN, CNN, LSTM, GRU	Acc, AUC, F1-Score, F2-Score, Sensitivity, Specificity

FIN-financial data, LING-linguistic data, TXT-text data, BBN-Bayesian belief network, NB-Naive Bayer, DTNB-NB with the induction of decision table, CART-classification and regression tree, LMT-logistic model trees, MLP-multi-layer network, GP-genetic programming, LR-logistic regression, DNN-deep neural network, PNN-probabilistic neural network, RF-random forest, SVM-support vector machine, Acc-Accuracy, AUC-area under the ROC curve, MC-misclassification cost, TPR-true positive rate, TNR-true negative rate, FPR-false positive rate, FNR-false negative rate, SFL-system function linguistics theory.

positive rate, TNR-true negative rate, FPR-false positive rate, FNR-false negative rate, SFL-system function linguistics theory.

Deep learning has getting a lot of attention lately with breakthroughs in many fields because of its strong learning ability. Textual analysis models based on DL can extract characteristics and distributed representation of data using multiple hierarchical structure. Hence, it is regarded as a promising solution for extraction of contextual information from document. Still, its application to fraud detection has not yet been explored deeply. In this paper, we are trying to apply deep learning techniques into text analysis in MD&A of listed company's annual report, and compare the performances with other traditional classification techniques used in financial statements fraud detection.

In addition, accuracy (Acc) and area under the ROC curve (AUC) are often used to measure the ability to distinguish fraud cases. Some studies also considered precision and recall as performance evaluation. In this paper, we provide a comprehensive evaluation of different deep learning techniques using some metrics, including AUC, sensitivity, specificity, F1-score, F2-score and accuracy.

III. RESEARCH METHODOLOGY

In our study, the objective is to present novel fraud detection models combined with the numeric and textual financial data as input from the Chinese listed companies' annual reports. For the reason that unstructured text cannot be directly used as input, they have to be transformed into numeric format, facilitating algorithmic processing without losing their information and content. Therefore, the following subsection first classifies Chinese words embedding, and then presents the most representative tools in deep learning for financial fraud detection in this paper.

A. CHINESE TEXT EMBEDDING

Normally, words in English can be easily recognized since the space token is a good approximation of a word divider. Different from English (or more broadly, languages that use some form of the Latin alphabet), there are no interval marks between words in Chinese (or other languages that do not have obvious word delimiters such as Korean and Japanese) [38]. Therefore, it is difficult for word segmentation to identify ambiguous words in Chinese document preprocessing. In order to transform the textual information in the listed companies' annual reports into numeric

vectors, there are two necessary steps: Chinese word segmentation (CWS) and word vector calculation. After that, the textual information will be fed as input for predictive modeling in deep learning.

Jieba is the most widely-used open-sourced Chinese word segmentation system up to now because of its excellent mapping ability [39]. More specifically, it provides interface for Python programming language, and the algorithm using Jieba is simple with high accuracy. After words segmentation, stop words are removed according to the stop words list, which is mainly generated for financial materials. Then, how to represent them with numeric vectors is crucial before feeding them as the input of predictive model. Frequency-based BOW (Bag of Word) embedding and prediction-based neural embedding are popular methods for text representation. However, the number of unique words in the document usually accounts for only a small part in the whole corpus, usually causing sparse vector for the document. Word2vec, proposed by Google in 2013, is an effective model used for creating lower dimensional and dense embedding for textual data [40]. There are two learning models in Word2vec: continuous bag of words (CBOW) and skip-gram [41], [42]. In detail, CBOW predicts a target w using n-length words before or after w, representing by $P(w|context)$, while skip-gram uses each word to predict the probability of its context, representing by $P(context|w)$. And the input is an initial word vector constructed according to the dictionary, and the output is the word vector of the predicted word.

B. DEEP LEARNING

Traditional machine learning algorithms have been extensively discussed and analyzed for financial detection in previous studies [43], [44]. These approaches are not very suitable for large dataset, particularly in handling of Chinese textual data. Recently, deep learning techniques, including convolution neural network (CNN) and Recurrent Neural Network (RNN), have been applied to many branches of engineering and sciences fields with large amount of data [45].

1) CONVOLUTIONAL NEURAL NETWORK

Convolutional neural network (CNN) is a type of neural network with short-term memory ability by means of a feature extractor composed of convolution layer and sub sampling layer. The feature maps, channels, pooling, stride and padding are the key terms in CNN [46]. In CNN, current output of a sequence is also related to the previous output because of such mechanism: it retains the previous information, and used as input to current output. Essentially, CNN is an input-output mapping, which can learn relationships between them without any accurate mathematical expression. Once the convolution network is trained with a known pattern, the network has the mapping ability between input and output pairs.

2) LONG SHORT TERM MEMORY NETWORK

From the network structure view, the recurrent neural network (RNN) will remember the previous information and use the previous information to generate the output of later nodes. In other words, the nodes between the hidden layers of the recurrent neural network are connected. And the input of the hidden layer includes not only the output of the input layer, but also the output of the hidden layer at the previous time [47].

The structure of long short term memory network (LSTM) is just like that of RNN by having a cell state with the memory of the network. The gates used in LSTM are the forget gate f_t, input gate i_t, output gate o_t, and input modulation gate $\hat{c}_t$. Forget gate f_t is used to decide which characteristics are extracted to calculate; input gate i_t determines whether there will be information input to the memory cell at this time and the output gate o_t decides the output information from memory cell. The interaction among the gates is noted in equations below, where $\odot$ represents element-wise multiplication.

$$
\begin{aligned}
i_t &= \sigma(U_i x_t + W_i h_{t-1} + b_i), \\
o_t &= \sigma(U_o x_t + W_o h_{t-1} + b), \\
f_t &= \sigma(U_f x_t + W_f h_{t-1} + b_f), \\
c_t &= f_t \odot c_{t-1} + i_t \odot \hat{c}_t, \\
\hat{c}_t &= tanh(U_c x_t + W_c h_{t-1} + b_c), \\
h_t &= tanh(c_t) \odot o_{t,}.
\end{aligned}
$$

In LSTM, the hidden state is obtained with a cell state passing through a neuron and an output gate. Therefore, the memory contains in hidden state is actually the content after attenuation of cell state. In this way, what stored in hidden state is mainly ''short memory'', while those stored in cell state is mainly ''long-term memory''. The existence of cell state enables LSTM to well characterize long dependency. Fig. 1 describes the complete flow of an LSTM cell, where each dotted box represents a single step [48].

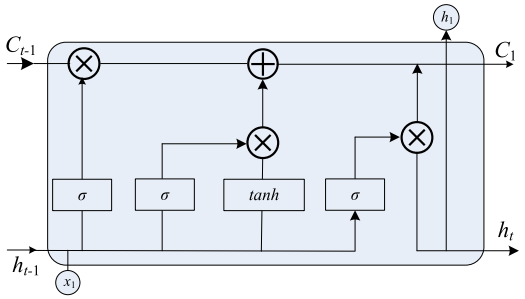

FIGURE 1. Structure of an LSTM cell.

Gated Recurrent Unit (GRU) network is the variant of LSTM,whose structure is shown in Fig. 2 [49]. Its main object is to reduce the gradient disappearance problem while retaining the long-term sequence information. Reset gate r_t and update gate z_t are two gates in GRU, where the reset gate r_t determines how to combine the new input information with the previous memory, and the update gate z_t defines the amount of previous memory saved to the current step.

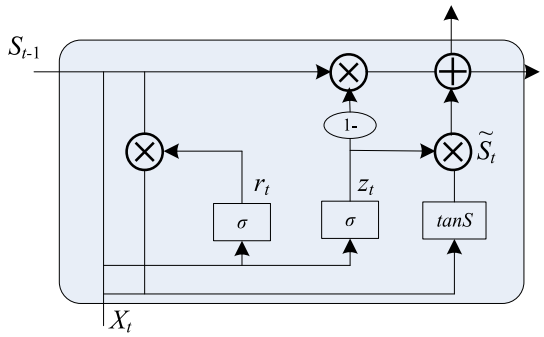

FIGURE 2. Structure of GRU model.

The forward formula of GRU can be obtained as follows.

$$\begin{aligned} r_t &= \sigma(W_r \cdot [S_{t-1}, X_t]); \\ z_t &= \sigma(W_z \cdot [S_{t-1}, X_t]); \\ \hat{h}_t &= tanS(W_h \cdot [r_t \odot S_{t-1}, X_t]); \\ h_t &= (1 - z_t) \odot S_{t-1} + z_t \odot \hat{h}_t; \\ y_t &= \sigma(W_0 \cdot h_t). \end{aligned}$$

There is no much difference between the LSTM and GRU. In LSTM, the new input is composed of the input of the current time and the output of the historical unit. However, the new input of GRU is composed of the input at the current time and the filtered historical unit output. Filter mechanism is not included in LSTM (but it can be regarded as that LSTM has been filtered at the time of output, that is, the output gate in LSTM can be regarded as the reset gate in GRU).

3) EVALUATION METRICS

There are multiple metrics used measuring the performance of any binary classification algorithms. Similarly, the financial statement fraud detection is typically regarded as a binary classification problem with four potential classification outcomes:

(i) True positive (TP): it denotes prediction results of those fraudulent companies are correct;

(ii) False negative (FN): it denotes prediction results of those fraudulent companies are incorrect, classifying them as non-fraudulent companies;

(iii) True negative (TN): it denotes prediction results of those non-fraudulent companies are correct;

(iv) False positive (FP): it denotes prediction results of those non-fraudulent companies are incorrect, classifying them as fraudulent companies;

Traditionally, the accuracy is widely used in model predictive power comparisons, which is defined as the percentage of correctly classified instances and all cases:

$$\text{Accuracy} = \frac{TP + TN}{P + N} \tag{1}$$

However, accuracy is not a suitable metric for fraud detection model evaluation due to the existence of high class imbalance problem in the datasets. To estimate the predictive power, many previous studies considered a combination of measures like precisions, sensitivity (also called TP rate or recall), *et al.* Nevertheless, model effectiveness measurement should pay much attention on high sensitivity by correctly classifying as many positive samples as possible. Therefore, model performance is evaluated by the AUC (Area Under Curve), sensitivity, specificity, F1-score, F2-score and accuracy in this study.

Receiver operating characteristic curve (ROC) shows the relationship between sensitivity and specificity by plotting the rate of true positives (fraudulent classified as fraudulent) to the rate of false positives (non-fraudulent classified as fraudulent). AUC represents the area under the ROC, whose values range from 0.5 to 1.0, and the higher the AUC, the better the model can distinguish between fraudulent and non-fraudulent cases. So, this study also employs the AUC as a measure of separability to compare the predictive performance of the models and determine their suitability.

The sensitivity represents the ratio between the number of correctly classified fraudulent instances and all fraudulent instances:

$$\text{Sensitivity} = \frac{TP}{P} = 1 - \text{FN rate} \tag{2}$$

The specificity describes the ratio between the number of correctly classified non-fraudulent instances and all non-fraudulent instances:

$$\text{Specificity} = \frac{TN}{N} = 1 - \text{FP rate} \tag{3}$$

The F-score is a combination of precision and sensitivity, which is used to measure how precise and how robust the models classify fraudulent cases:

$$F_\beta - \text{score} = (1 + \beta^2) \times \frac{\text{precision} \times sensitivity}{\left(\beta^2 + \text{precision}\right) + sensitivity} \tag{4}$$

IV. DATA

Despite the existing guidelines, the fraud detection can be one of most challenging and important task with highly imbalanced dataset because ratio of non-fraud listed companies is very high compared to fraudulent companies. For example, the proportion of statements that were fraudulent and non-fraudulent in the annual reports submitted to the China Securities Regulatory Commission (CSRC) for the period from 2016-2020 was approximately equal to 1:20. In the previous researches, the number of companies that committed fraud is different from tens to thousands. In addition, most current studies adopted an approach by matching the non-fraudulent companies with the fraudulent companies using metrics such as year, scale and industry.

A. LABELLING

In this study, all the data are from China Stock Market & Accounting Research Database, accessing by: https://www.gtarsc.com/, where listed companies' annual financial reports are publicly available. Also, the financial reports disclosed as fraudulent are identified by the China

Securities Regulatory Commission (CSRC) during period of 2016-2020. The dataset provides the extremely rich information resources that cover all aspects of companies, not only values of financial variables but also textual analysis in the annual reports.

Initially, our sample contained data from 1068 distinct Chinese listed companies on the Shenzhen Stock Exchange between 2016 and 2020 (There were some de-listings of Chinese companies during this period). And we have analyzed the number of companies per sector in Table 2, after excluding the sectors of banking, utilities and financial services from the samples. Ultimately, the resulting dataset contains 244 instances of companies annual reports that committed fraud and 4886 instances of companies annual reports that had not committed fraud, showing the typical class imbalance problem. In other words, 5130 annual reports constitute the dataset as the input to build classification model in this paper, which will predict whether a company is likely to commit financial fraud.

TABLE 2. The number of listed companies per sector.

Sector	Number of Companies
Industrial Goods & Services	646
Wholesale and retail	105
Information technology	87
Construction & materials	50
Health care	6
Education	14
Financial service	70
Accommodation & catering	18
Water & power	20
Science	16
Comprehensive	36
Total	1068

The dataset consists 5130 Chinese A-share listed companies' annual reports from 2016-2020, and all the records are labeled with ST (special treat) or normal. In general, a company is marked with ST when heavily getting into serious financial crisis. As a general rule, a company is marked with ST because of the following reasons: (i) there are two consecutive annual losses after audited by accounting companies; (ii) the net income per share of listed companies is lower than its face value; (iii) both of (i) and (ii) are satisfied. Therefore, we find 240 distinct fraudulent companies with 244 fraud-year samples during this period. And the rest of 4886 samples are marked with non-fraud-year samples in this research. Table 3 presents the number of normal and ST marked listed companies, respectively.

In addition, the deadline of a listed company releasing its own financial reports is before April 30 each year according to China Annual Report Disclosure System of List Companies. Once a company is marked with ST, it refers to the previous financial conclusion, not the data of current year. As can be seen from Table 3, the final dataset is reduced to 5130 company-year observations. Furthermore, we perform the extraction of numeric and textual data from the annual reports that may cover management business of listed companies.

TABLE 3. The number of normal and ST marked listed companies.

Item	Number of Training	Number of Testing	Total
Number of ST Companies	206	38	244
Number of Normal Companies	4126	760	4886
Total	4332	798	5130

B. TEXTUAL DATA IN ANUUAL REPORT

Listed company's annual report is the main content of mandatory and regular information disclosure in Chinese stock market. Also, it is the primary approach to well understood its real financial situation and the future trend for potential investors, auditing companies and state regulators. In recent years, more and more researchers have begun to pay much attention to management and comment information disclosure in the annual financial statement. And the existing studies have shown that the non-numerical information disclosure could be helpful in evaluation of a company value, companies' cost of capital decrease, minimum of analysts' expectations error and improvement of audit quality [43]–[46].

The essential section, commonly called ''Management's Discussion and Analysis of Financial Condition and Results of Operation (MD&A)'' in listed company's annual report, is a useful, necessary and indispensable supplement for investors to grasp the future development direction of the company. This part also offers the analysis of important events, trends and uncertainties that will affect the future of the company. In 2002, the securities supervising administrative department in China brought in this system so that the sponsor can continuously instruct and supervise the issuing listed company, and finally enhance the quality of listed company and protect the interest of the investors. More recently, some researchers have emphasized the increasing significance of textual analysis of financial documentation. Deep learning approach is well suit for the textual analysis of MD&A section because nearly all the MD&As have the same structure. Before 2015, MD&A had always been the main content of the chapter ''report of the board of directors''. And it was made as an independent part (the 4th section) in the financial report since 2016. And then its title was modified to ''Business Situation Discussion & Analysis (BSD&A)'', including introduction, main business analysis, balance sheets analysis, the core competences analysis, investment analysis and future developing prospect.

In the rest of this paper, we still use MD&A denoting this part in order not to cause confusion.

In this study, 1068 Chinese listed companies' annual reports are collected and pre-processed using Chinese text mining pattern, constituting the primary source of raw text data. We first employs character-based Chinese morphological analysis for segmenting Chinese texts into words, and then presents a method based on structure information of constituent characters. The text length of each segment was limited to 2000 words, and the part in excess of the text shall be cut down from the tail, for the reason that more important sentences are usually placed on the front according to Chinese writing styles.

C. QUANTITATIVE DATA IN FINANCIAL STATEMENT

Along with text features, quantitative financial variables are particularly important, which can clearly obtain the operating status and the performance of the firms. And the existing studies have shown the relationship between the quantitative data and frauds from financial statements. Following the guidelines of existing researches, the financial and non-financial variables are extracted from listed companies' annual reports (described in the next section). Specially, financial variables include indicators like total assets (adopted as a proxy for company size), activity ratios, solvency and inventories as non-cash working capital drivers [35], [36].

V. FINNCIAL AND NON-FINANCIAL INDICATORS SELECTION

As is shown that there are no fixed indicators used as signs of financial statements detection for the reason that financial indicators mainly reflect the listed company's financial situation only from one of its aspects, failure to comprehensively address the problem of its management and future development. Until now, there are no consensus on what best variables group is in financial fraud detection. Nevertheless, previous literature has proved the importance of financial variables, and have proposed a number of ratios in the past years [50]–[52]. After reviewing several existing studies and governance structure section on the CSMAR database related to fraud detection, this paper proposes two types of indicators: financial indicators and non-financial indicators.

A. FINANCIAL INDICATORS

It is found that some financial variables are more important than others for the prediction purpose, whereas some have negative impacts on the classification accuracy. Financial indicators selection is regard as paramount importance for any learning algorithm and usually leads to problems related to incomplete or irrelevant information when poorly done. Therefore, appropriate variables should be selected with the purpose of identifying the statements with financial fraud.

Combined with existed research results, this study proposes financial and non-financial variables referenced from methods of machine learning techniques. And the first-level input financial variables can be divided into ten categories: solvency, activity, profitability, EVA (Economic value added), liquidity, development capability, risk level, structure ratio, index of per share and market value. A total of 58 financial indicators obtained as a result constitutes the corresponding second-level variables in this research, as appear in Table 4.

Most of the selected factors are consistent with prior studies. And the top fraud factors are described as log of total debt, equity, debt to equity, total assets, net fixed assets to total assets. In addition, the profitability, liquidity, solvency, activity and structure ratios are significant predictors for fraud detection.

Additionally, most of the current studies focused on the financial variables, overlooking the non-financial factors in financial fraud detection, which are related to the corporate governance structure, thus affecting its operational status. In order to obtain the fraud clues, much attention should be paid to those non-financial factors, such as ownership structure, management structure and auditor's opinion. And the non-financial variables employed in this study can be seen in Table 5.

The original data may be mixed with noise, distortion or extreme values, and need to be properly preprocessed, involving several steps, including cleaning and normalizing the raw data before being used for modeling. To solve missing values problem, we have removed those samples from the dataset if the value of one or more attributes be missed. Beyond that, the variables in dataset are not in the same scale, usually result to poor performance when fed as input to deep learning models. Scaling and standardization methods bring the features together to almost the same scale, which make them more suitable for algorithm input. Also, as the data volume collected in this study is large, and the financial items in the reports have a wide range, it is necessary to reduce data dimension for ensuring the accuracy and reliability of the data analysis and mining results.

According to the variables in Table 4 and Table 5, $x_{ij}, i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$, denotes the m^{th} listed company's n^{th} feature if the number of listed company is m and each has n features for analysis. The matrix can be represented as follows.

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{bmatrix}$$

Each element in matrix X is transformed to $\hat{X}$ using mean normalization, and the value of $\hat{x}_{ij}$ is calculated as follows.

$$\hat{x}_{ij} = x_{ij} - \mu(x_j), \tag{5}$$

where $\mu(x_j)$ is the average value of column j.

After that, covariance matrix D of $\hat{X}$ can be obtained using the following formula:

$$D = \frac{1}{m-1}\hat{X}^T\hat{X}, \tag{6}$$

where $\hat{X}^T$ is the transpose matrix of X^T.

TABLE 4. Financial variables of listed companies for fraud detection.

Dimension	Second-level variables	Dimension	Second-level variables
Solvency	X1: Current ratio	EVA	X30: The sale profit ratio
	X2: Quick ratio		X31: Working capital turnover ratio
	X3: Total debt	Liquidity	X32: Net Earnings and cash flows
	X4: The logarithm of total debt		X33: Working capital
	X5: Equity		X34: Working Capital/total assets
	X6: Debt to equity		X35: Current assets/current liabilities
	X7: Total debt/total assets		X36: Current assets/total assets
	X8: Long term debt/total assets		X37: Cash and deposits/total assets
	X9: Short term debt/total assets		X38: Quick assets/current liabilities
Activity	X10: Account receivable/sales		X39: Cash flow/total debt
	X11: Inventory/sales		X40: Cash flow/current debt
	X12: Inventory/total assets		X41: Cash flow/cash dividend
	X13: Inventory turnover		X42: Cash flow/equity
	X14: Sales growth	Development	X43: Capital accumulation ratio
	X15: Sales		X44: Asset inflation and incremental ratio
	X16: Inventory		X45: Revenue growth rate
Profitability	X17: Gross margin		X46: Profit growth rate
	X18: Assets return ratio		X47: Total assets growth rate
	X19: Net margins of total assets	Risk	X48: Financial leverage
	X20: Gross profit/total assets		X49: Operating leverage
	X21: Net profit/total assets	Structure	X50: Liquidity ratios
	X22: Net profit/sales		X51: Cash asset ratio
	X23: Net income/fixed assets		X52: Equity/total assets
	X24: Earnings before interest and taxes	per share	X53: Cash dividends per share
	X25: Ebit/total assets		X54: Total income per share
	X26: Z-score		X55: Revenues per share
	X27: Net profit after tax	Market value	X56: Price-earnings ratio
	X28: Cash and deposit/current assets		X57: Price-book ratio
	X29: Return on invested capital		X58: Tangible assets ratio

After calculating the eigenvalues of D and sorting them with descending order, the top-k data constitutes the matrix P. And the lower-dimensional matrix can be obtained using the following formula.

$$Y = P\hat{X}. \tag{7}$$

Based on these, each data in Y is handled with maximum and minimum normalization using following formula.

$$Y'_{ij} = \frac{Y_{ij} - \min\{Y_j\}}{\max\{Y_j\} - \min\{Y_j\}}, \tag{8}$$

where Y'_{ij} is the normalized value, y_{ij} is the element in i^{th} row and j^{th} column; $\{Y_j\}$ refers to the whole data in j^{th} column.

Thus, the final datasets can be obtained to build a classification model, predicting whether there is the possibility of financial fraud with listed company's annual report.

VI. CLASSIFICATION RESULTS AND ANALYSIS

The main objective of this paper is to compare the fraud predictive capacities with numeric and textual data in Chinese listed companies' annual report using deep learning against traditional machine approaches. In this part, we will present the classification results by means of empirical analysis using a set of machine learning models, including random forest, SVM, XGB (eXtreme Gradient Boosting), ANN and deep learning models, such as CNN, LSTM, GRU and transformer. These models generate fraud classification results based on financial variables, non-financial variables and the text features extracted from Business Situation Discussion & Analysis (BSDA) in listed companies' annual reports.

Fig. 3 shows a complete experiments workflow designed to test all the predictive models with different learning techniques. First step of the framework is data collection, and all the data related to the listed companies have been collected from their annual reports, including not only financial statements namely balance sheet, income statement and cash flow statement, but also management analysis data in annual reports. After that, data preprocess is necessary for the purpose of cleaning the noisy and wrong data. Also, data

TABLE 5. Non-financial variables of listed companies for fraud detection.

Dimension	Second-level indices	Dimension	Second-level indices
Ownership structure	Y1: Ownership concentration index CR1	Management structure	Y9: The number of employees
	Y2: Ownership concentration index CR5		Y10: Supervisors' size
	Y3: Index *Z*		Y11: The number of senior supervisors
	Y4: Index *S*		Y12: Total annual salary of director, supervisors and senior supervisor
	Y5: Relations among top 10 shareholder		Y13: Total annual salary of top 3 director, supervisors and senior supervisor
Management structure	Y6: Chairperson and general manager holding a concurrent post		Y14: Total annual salary of top 3 directors
	Y7: Number of directors (chairman included)		Y15: Total annual salary of top 3 senior supervisors
	Y8: The proportion of the independent board member	Audit opinion	Y16：A standard and unqualified auditor's report

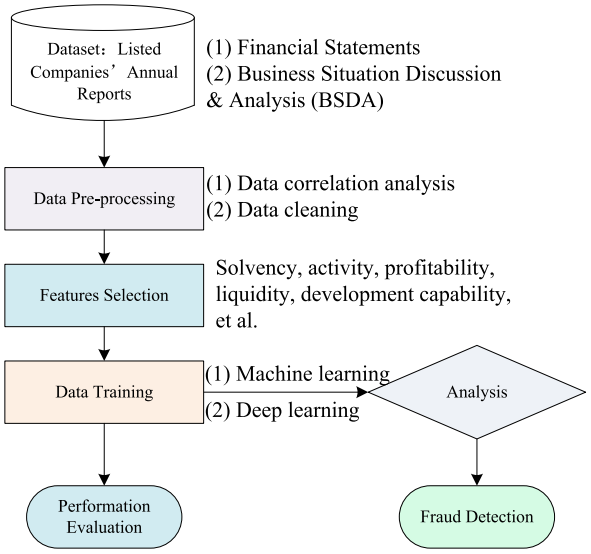

FIGURE 3. Financial fraud detection workflow model.

cleaning, data transformation, data integration and data reduction are included in this phase in order to prevent data inconsistency. The framework suggests the combination of digital and textual data for detection of financial statement fraud. The main function of feature selection is to reduce the computational overhead and improve the classification performance. It also eliminates irrelevant or redundant features, so as to reduce the number of features, improving the accuracy of the model and reducing the running time. After that, the dataset was classified randomly as a training set (70%) and a test set (30%). Then, processed data is fed as input to the algorithms and output is obtained. Finally, the simulation outcomes will be evaluated and compared with the traditional classification techniques in terms of AUC, sensitivity, F-score and classification accuracy.

A. MODELING OF NUMERIC FINANCIAL DATA

Modeling using the quantitative data from financial statement has always been one of the most common methods in financial statement fraud detection. In the experiments, models such as RF, SVM, XGB, ANN, CNN and XGBoost are employed in order to explore performance improvement ways. In addition, trained models were evaluated by calculating metrics, such as AUC, recall, precision and F-score, using Python's pandas and Sklearn library.

In the first experiment, we have not used the features selection, but fed the original dataset as the input fed to the algorithms. Table 6 summaries the results of the experiment using all financial variables.

The results in Table 6 show that CNN and RF performed much better than the other methods in terms of most classification metrics. On the one hand, the CNN achieved the best results in terms of AUC, specificity, F1-Score and F2-Score, indicating its well predictive performance in fraud detection. On the other hand, the RF achieved the optimum performance in terms of sensitivity and accuracy. In contrast, LR performed poorly in the classification of the fraudulent classes. In addition, these results are in line with Liu *et al.* [53], who showed that RF model always performed especially well with multivariable applications, such as financial statements, because it can easily deal with high-dimensional dataset with their feature subsets random selection. Also, the results are similar to Hajeck and Henriques' in reference [26], where an accuracy of 88.1 percentage on financial data with algorithm C4.5 was reported.

In the second experiment, we used the selected financial variables, including financial and non-financial variables of listed companies, these variables have been shown in Table 4 and Table 5 in Section V. And Table 7 reports corresponding results of performance metrics for each selected model.

TABLE 6. Classification results with different models using all financial variables.

Classifier	AUC	Sensitivity	Specificity	F1-Score	F2-Score	Accuracy
LR	0.7251	0.6925	0.7591	0.4873	0.7594	0.8109
RF	0.8106	**0.7748**	0.8006	0.5862	0.7922	**0.8774**
SVM	0.7754	0.5932	0.7765	0.4781	0.7741	0.8392
XGB	0.8096	0.6963	0.8695	0.6132	0.8263	0.8418
ANN	0.7364	0.7654	0.7125	0.5451	0.7125	0.7456
CNN	**0.8851**	0.7852	**0.8936**	**0.6057**	**0.8369**	0.8419

TABLE 7. Classification results with different models using selected financial variables.

Classifier	AUC	Sensitivity	Specificity	F1-Score	F2-Score	Accuracy
LR	0.7816	0.7024	0.7683	0.4804	0.7624	0.8351
RF	0.8812	**0.7864**	0.8012	0.5603	0.8091	**0.8784**
SVM	0.7831	0.6288	0.7808	0.4836	0.7834	0.8681
XGB	0.8235	0.6924	0.8908	0.6067	0.8394	0.8472
ANN	0.7654	0.7684	0.7002	0.5091	0.7035	0.7012
CNN	**0.8920**	0.7946	**0.9082**	**0.6094**	**0.8464**	0.8647

As can be seen from Table 7, the classification results using selected financial variables have a slightly improvement compared to the results with all variables, though some results show statistically similar. In other words, omitting part variables from the set of original dataset have not brought about negative influence on the prediction ability of the models. In addition, the RF and CNN still have better performance at predicting fraud on financial statement in terms of AUC and accuracy, showing a non-linear dependency between financial variables and the frauds status of listed companies' annual reports. Furthermore, the results in this study is consistent with the result of Kim *et al.* in reference [48], offering the SVM as high accurate. SVM is considered by many researchers to be a very effective classification method, and has been applied in fraud detection and other fields [54]–[56]. Also, it should be noted that ANN shows less impressive predictive performance, but obtaining satisfactory effect in terms of specificity. Compared with artificial neural network, CNN represents a promising classifier in ML, its high performance is noteworthy since it was not considered in previous work on fraud detection. And CNN gives the best prediction performance in terms of AUC, F1-Score and F2-Score. Considering that the cost of missing an actual fraud case is much higher than a false alarm, this study shows that the F2-Score is the most suitable threshold-based indicator of model performance. Based on the above analysis, deep learning model, such as CNN, is a feasible solution when the aim is to identify those listed companies with numeric data in annual reports.

B. MODELING OF TEXTUAL DATA

As the first step towards fraud prediction from Chinese textual data, four DL methods are employed in textual data mining, namely RNN, LSTM, GRU and Transformer (based solely on attention mechanisms). In addition, RNN, LSTM and GRU in particular, have been shown as the most advanced approaches in sequence modeling and transformation problems. Transformer is a model architecture eschewing recurrence with transformer blocks as feature extractor. Also, we compare their performances with traditional models, for example, Goel *et al.* [57] utilized SVM as classifiers and achieved accuracies of 0.8950 using the BOW model to perform modeling on text data.

In the aspect of text pretreatment, this paper contains two approaches: Chinese word segmentation and embedding processing based on word vector. The automatic word segmentation of Chinese sentences is still challenging when the unrestricted texts processing in annual reports are large. Due to the nature of specific domain, most of existing segmentation tools cannot achieve appropriate segmentation results. In this way, we selected the Jieba package when processing words segmentation [58], which is regard as the best Python Chinese word segmentation module. Specifically, Jieba supports four segmentation methods: accurate mode, full mode, search engine mode and paddle mode, providing a great help for Chinese natural language processing in this paper. After that, it is necessary to represent the words or phrases with values before fed them to the algorithm. And the word vector model is designed by Institute of Chinese Scientific Space (https://spaces.ac.cn/archives/4304) in our study, whose word vector library contains more than 400,000 entries ensuring the embedding effect. Besides that, its training tool is Word2Vec in Gensim with 8,000,000 articles on Wechat as corpus.

In optimizing the control parameters, we directly take Focal Loss as their loss function as a result of samples imbalance problem. Also, we observe that the improvement of loss function can make sure the data feature learning efficiency and effectiveness, avoiding all the companies marked

with non-fraudulent during the initial period. As shown by statistical data, the text length of MD&A in most annual report exceeds 1000, for example, the number of words in 2020 Kweichow Moutai annual report is 1587. Even if some infrequent and obscure words were removed, the segmentation results is still more than 800. In this way, we set the maximum text length to 600 and 1000 separately. To solve this problem, some text in managerial comments of annual report might be truncated while those insufficient text will be replenished repeatedly.

Table 8 and Table 9 present overviews of the modeling results with the predictive methods for textual data in listed companies' annual reports with 600-word and 1000-word, respectively.

As can be seen from Table 8, deep learning methods perform better compared to the machine models in terms of classification accuracy, F1-score, F2-score, AUC as well as specificity. However, the performance using textual data mining performed worse compared to digital financial variables, which indicates that the prediction fraud only using the textual variables has little significance in practice regardless of the classifier.

Moreover, we can observe that modeling text with 600-dimension words vectors provides approximate results across all models in comparison to financial data, which falls in line with the previous work, such as reference [57] and references [59], [60]. And this reflects the relationship between the richer content of ''Business Situation Discussion & Analysis (BSDA)'' and the financial data in annual reports.

From the results in Table 9, we can see that the performance using 1000 words have slightly improved compared to the results using 600 words. And RNN based models exhibit better performance compared to those models based on CNN and Transformer. Be noted that, the loss function in our experiments is replaced with focal loss function, whose definition is as follows:

$$\mathrm{FL}(p_t) = -(1-p)^{\gamma}\log(p_t), \tag{9}$$

where a modulating factor $(1-p)^{\gamma}$ to the standard cross entropy loss is added, with tunable focusing parameter $\gamma \geq 0$. It has been shown that the loss function has the advantage of previous approaches when dealing with class imbalance problem [61]. And the same results can be achieved as in Table 4 that most of the architectures seem to benefit the most in terms of accuracy.

Meanwhile, the length of input text in MD&A is another important factor on prediction performance. And it is observed that the predictive power of 1000-word input is superior to the 600-word with nearly all metrics, and is more reliable with RNN based models. Specifically, some important information will be lost with high probability when part of the text is truncated. On the other hand, the longer input text will have to increase the difficulty of model learning greatly: too much texts contain more complex information with much difficulty in understanding. In addition, previous information will be lost during recursion with too long text input, unable to handle to total valid information. However, it is essential to note that, the predictive performance will fall subsequently with the increased text length of input. And the maximum with 1000 words input get the best performance, indicating over-fitting problem.

Additionally, all deep learning models are employed by bidirectional recurrent neural network, implemented by bidirectional decorators in Keras. And this approach has been widely adopted in text classification application, for the reason that we usually placed the important information at the beginning and end of the paragraph, while the repetitive or secondary information at the middle part. As is known from the structure of bidirectional recurrent neural network, the earlier information memorized will gradually decrease during each iteration of updating the information inside the neuron. Therefore, forward recurrent neural network will pay more attention to the end of text, while backward propagation neural network focuses on the beginning of text. In this way, bidirectional recurrent neural network combined the two types as measures to ensure the consideration of text beginning and end.

Deep learning models exhibit improved performance in comparison to the traditional machine learning setup, especially GRU(256) and Transformer. Comparative analysis revealed that the LSTM with 128 neurons show superior accuracy compared with that of 256. However, GRU with 256 neurons performs better than that of 128 neurons. The possible reasons for the result are that GRU architecture simplifies the gate structure that controls the network memory and improves the performance dealing with long sequence input. While the memory information structure in LSTM is more complex, resulting valid information loss processing long text and ultimately leading to lower model performance. Nevertheless, RNN is still regard as an advanced technology in natural language processing. And its performance can be explained by the intrinsic capacity to extract significant contextual similarities within documents.

Based on the analysis above, it is necessary to combine both of financial numeric and textual variables as the input in the prediction of financial frauds, which indeed is discussed in the next sub-section.

C. MODELING OF COMBINATION OF FINANCIAL NUMERIC DATA AND TEXT DATA

Although the textual data in MD&A of annual report have exhibited predictive power in financial fraudulent using deep learning methods. However, the decrease is also visible for the metrics, such as recall and precision, leading to the conclusion that only text data cannot provide better predictive performance than financial digital data. From this point of view, it is of little or no use to construct a financial fraud model only with text solely as input. Therefore, the input setup with a combination of words vectors and digital financial data is still at the core of our study. In this way, a combination of word vectors with financial variables into one dataset

TABLE 8. Classification results with different models for text data with 600-word.

Classifier	AUC	Sensitivity	Specificity	F1-Score	F2-Score	Accuracy
SVM	0.6123	0.6004	0.7245	0.4128	0.7125	0.7109
XGB	0.6584	0.6122	0.7341	0.5147	0.7249	0.7006
ANN	0.7006	0.6832	0.7429	0.4682	0.7136	0.7331
CNN	0.8021	0.7009	0.7940	0.5739	0.7364	0.7425
LSTM(128)	0.8351	0.7154	0.8001	0.6034	0.7458	0.7354
LSTM(256)	0.8465	0.7266	0.7946	**0.6551**	0.7631	0.7369
GRU(128)	0.8157	0.7241	0.8001	0.6187	0.7862	0.7547
GRU(256)	**0.8584**	0.7457	**0.8124**	0.6544	0.7952	0.7538
Transformer	0.8321	**0.7654**	0.8117	0.6420	**0.8012**	**0.7855**

TABLE 9. Classification results with different models for text data with 1000-word.

Classifier	AUC	Sensitivity	Specificity	F1-Score	F2-Score	Accuracy
SVM	0.6846	0.6154	0.7261	0.5124	0.7296	0.7315
XGB	0.7015	0.6351	0.7352	0.5241	0.7364	0.7415
ANN	0.7674	0.7141	0.7424	0.4957	0.7347	0.7345
CNN	0.8125	0.7251	0.8014	0.5748	0.7651	0.7498
LSTM(128)	0.8241	0.7554	0.8136	0.5941	0.7716	0.7598
LSTM(256)	0.8398	0.8014	0.8247	0.6049	0.7808	0.7629
GRU(128)	0.8256	0.8157	0.8198	0.6365	0.8041	0.7937
GRU(256)	0.8549	0.8396	0.8584	0.6541	**0.8165**	0.7848
Transformer	**0.8664**	**0.8487**	**0.8597**	**0.6874**	0.8084	**0.7921**

is used as the input in the prediction of financial fraudulence. Meanwhile, we try to produce an auxiliary classification result using deep neural network, and analyze synthetically with previous results.

Purda and Skillicorn [32] have attempted to do such a thing and conducted a comparison of text with financial data proposed by Dechow *et al.* [33] separately. And these two methods are complementary to each other because each of them only competes part of fraud detection [62], [63].

Table 10 and 11 report the corresponding classification results with different models for combination of financial variables and text data of 600-wrod and 1,000-word, respectively.

From Table 10, we can observe that the combination of FIN+TXT input setup exhibits improved performance in comparison to the financial or text data solely, especially those models of SVM and GRU(256). What is especially important is that F2-score increases across the ML benchmarks, which is different from F1-score. Also, we can see that GRU(256) offers the best performance with AUC 94.49% in Table 10, followed by LSTM(256) and Transformer with AUSs of 93.98% and 93.64%, respectively. And the results of modeling on LSTM and GRU show powerful predictive performance with diverse inputs, facilitating the further exploration of data enrichment for fraud detection. Considering the comparison among LSTM (256), GRU (256) and Transformer, it is shown that the performance of transformer has not significantly improved with FIN+TXT data across all metrics. And this indicates that latest technology does not necessarily lead to the superior application results. Similarly, as can be seen from Table 11 that the performance of combination input with 1,000 words has slightly improvement in comparison to input with 600 words in textual data, proved that the much text provides plentiful meaning in the fraud detection. Also, the deep learning models could correctly identify most frauds with higher metrics and offer substantial improvements for fraud detection, not only using the financial variables as input but also using the textual data as input. Besides that, both the deep learning methods and the traditional algorithms have achieved good results, and RNN based approaches have slightly better performance among tested algorithms. In addition, with the support of GPU computing, deep learning methods usually take less training time than those of traditional methods on large dataset, when implemented based on the tensorflow library.

The results of combination input allow us to conclude that the proposed DL models offer substantial improvements with different metrics for fraud detection facilitation. Additionally, the latest techniques can effectively reduce the misclassifications ratio and time consumption, and therefore improve financial statement detection efficiency and quality. And the DL capacity might be particularly important for practitioners, given the need to substantiate the audit judgement.

VII. DISCUSSION

We have explored the predictive capacities of deep learning with the Chinese listed companies' annual reports mining for

TABLE 10. Classification results with different models for combination of financial variables and 600-word text data.

Classifier	AUC	Sensitivity	Specificity	F1-Score	F2-Score	Accuracy
SVM	0.8536	0.6843	0.8261	0.5645	0.8254	0.8699
XGB	0.8751	0.7045	0.9087	0.6582	0.8669	0.8963
ANN	0.8067	0.7741	0.8824	0.6615	0.7916	0.7743
CNN	0.9071	0.8080	0.9234	0.7864	0.8608	0.8924
LSTM(128)	0.9218	0.8454	0.9236	0.8041	0.8216	0.9098
LSTM(256)	0.9398	**0.9014**	0.9147	0.8749	0.8408	0.9329
GRU(128)	0.9356	0.8857	0.9098	0.8865	0.8641	0.9237
GRU(256)	**0.9449**	0.8996	0.9184	0.8941	**0.9165**	**0.9448**
Transformer	0.9364	0.8887	**0.9297**	**0.9374**	0.9084	0.9421

TABLE 11. Classification results with different models for combination of financial variables and 1000-word text data.

Classifier	AUC	Sensitivity	Specificity	F1-Score	F2-Score	Accuracy
SVM	0.8902	0.7856	0.9012	0.65019	0.8582	0.8109
XGB	0.8836	0.6994	0.9041	0.7102	0.8784	0.9006
ANN	0.8806	0.7610	0.9113	0.6808	0.8947	0.9031
CNN	0.9312	0.8304	0.9351	0.8077	0.9029	0.9165
LSTM(128)	0.9344	0.8528	0.9346	0.8315	0.8307	0.9254
LSTM(256)	0.9367	**0.8921**	0.9324	0.8689	0.8505	**0.9498**
GRU(128)	0.9418	0.8471	0.9548	0.9326	**0.9418**	0.9317
GRU(256)	**0.9581**	0.8595	**0.9634**	**0.9345**	0.9287	0.9462
Transformer	0.9501	0.8854	0.9418	0.9236	0.9148	0.9445

fraud detection. Still, there are questions needed to explain, including Chinese textual data mining and imbalance data treatment technologies.

A. CHINESE TEXT MINING WITH DEEP LEARNING MODELS

As mentioned in the literature review, few researchers have attempted to combine financial numeric variables and Chinese textual data for financial statement fraud detection. The exceptions are the studies by Hajek and Henriques [26] and Throckmorton *et al.* [64]. Although they have taken advantage of both financial data and linguistic data when predicting financial statement fraud, they haven't combined them as input to the algorithms. Furthermore, fraudulent identification is unlike traditional sentiment and subject classification, which can classify only using few keywords and simple rules. In this way, the predictive power is a little weak at text classification compared to other classification models. Aimed to solve this problem, we have adopted text classification models based on DLs, such as CNN and RNN. Compared with the traditional approaches of vocabulary counting, deep learning neural networks can be able to extract local feature of text and learn their regular pattern, while statistic-based method only utilized the words frequency of text, omitting information such as the combination of words in context. As a result, we can see that the GRU with 256 neurons exhibits high predictive capability with AUC 95.81% in test dataset, making it a promising model for fraud detection.

It should be noted that the structure and content of listed company's financial statement in annual report is not exactly the same with different years. In this way, we constructed the purely financial pre-warning model with 10 dimensions 58 financial variables and 4 dimensions 16 non-financial variables. Then we fed them as input to the classifiers. Also, the results of the AUC shows that the financial variables extracted with our approach have increased significant value to fraud detection models in combination with textual features.

In addition, the heterogeneity in performance shifts among various data types for models, showing that different models pick up on different signals. And the supplement of text information improves the performance, offering the recall of 92.5% with GRU model, indicating that the combination of these information might be more effective than the use of single information source in the decision-making processes of stakeholders. Companies those predicted with suspected financial fraud can be further investigated and examined by handwork calculation to see if they really have any frauds. While the false negatives companies can bring in significant risks because there are no ways to find out again.

B. IMBALANCE DATA TREATMENT

Generally speaking, class imbalance usually appears when the number of different types of data labels varies greatly, thus the data is separated into majority and minority. When the dataset is highly imbalanced, less reliable performance of the model is obtained if there is no preprocessing,

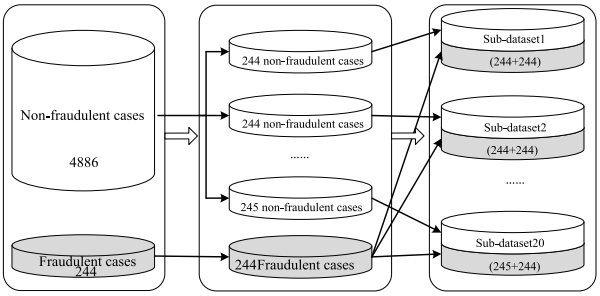

FIGURE 4. Data preprocessing to resolve class imbalance problem.

primarily because the minority are usually misclassified by treating them as noise [65]. Our dataset is also greatly imbalanced, since only a small number of companies commit financial statement fraud, and the majority companies in the dataset are non-fraudulent.

Previous studies have already recognized the complexity of the imbalance problem [66]–[68], and provided various solutions including over-sampling [69], under-sampling [26], [70]–[72] and synthetic minority over-sampling technique (SMOTE) [64]. However, the experimental results showed that most applications were not very effective for the reason of data preprocess measures. For example, over-sampling technique creates many extra copies, while under-sampling method has not used all available data instances when training, but only part of majority data. And SMOTE did not use the real data but synthetic data from minority.

Similarly, our imbalanced dataset consists of 244 fraudulent cases and 4886 non-fraudulent cases. For the purpose of the class imbalance problem solving, we follow the approach as in reference [35] and make 20 balanced sub-datasets as follows:

(i) We divide the whole dataset of non-fraudulent into 20 groups with 244 (or 245) cases;
(ii) We combine each non-fraudulent group with the 244 fraudulent cases to construct 20 balanced sub-datasets with 488 (or 489) cases;
(iii) We perform many experiments with the combined sub-datasets to calculate the more reliable results.

As can be seen in Fig. 4. And the final results have shown that overall performance is increased with the proposed sampling approach.

VIII. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

While financial fraud has a negative impact on economic and social development, it also causes huge losses to different stakeholders. However, detecting financial statement fraud is fairly challenging using traditional approaches due to companies' stratagem. Our main purpose of conducting this research is building models with high classification performance and deriving classification framework which can be used to detect the frauds with textual and numeric data in Chinese listed companies' annual reports. As the most advanced information processing technology, deep learning has made great achievements in many applications. In this way, this paper gives a framework for how this technique can be used in financial statement detection with Chinese companies' annual reports. Besides numerical data in financial statements, we analyze the ability of textual data attached to annual reports in financial statement fraud prediction and highlight the importance of textual analytics for detecting fraud with financial documents. Also, the results have shown that the deep learning models achieved considerable improvements in AUC compared to the earlier studies on the financial fraud detection. Furthermore, the textual information of the MD&A section of annual reports extracted through deep learning has the ability to improve the accuracy of financial statement fraud model detection, particularly in the highly unbalanced case of fraud detection.

In addition, there are some limitations in this research and can be extended in a few aspects. As the sampling period of the study is five years, some companies may have been delisted for some reasons. Also, there are some companies' annual reports have to be eliminated because of their incompleteness, which may affect the prediction results. Besides, the applicability of the models may need to further study, for the reason that the data source only involve Chinese listed companies, excluding those in other markets.

With regards to the future research directions, we can extend this work by extracting information from listed companies announcements. These can be helpful for easy understanding of prediction process because of rich seam of information. In addition, text mining algorithms for sentiment analysis of the textual description in financial statements can be optimized to provide better prediction performance of financial statement fraud.

ACKNOWLEDGMENT

The authors are very thankful to the China Stock Market and Accounting Research Database (CSMAR) for providing them with the dataset for conducting various experiments reported in this paper.

REFERENCES

[1] C. E. Hogan, Z. Rezaee, R. A. Riley, and U. K. Velury, "Financial statement fraud: Insights from the academic literature," *AUDITING, J. Pract. Theory*, vol. 27, no. 2, pp. 231–252, Nov. 2008, doi: 10.2308/aud.2008.27.2.231.
[2] Q. Chen, K. Kelly, and S. E. Salterio, "Do changes in audit actions and attitudes consistent with increased auditor scepticism deter aggressive earnings management? An experimental investigation," *Accounting, Org. Soc.*, vol. 37, no. 2, pp. 95–115, Feb. 2012, doi: 10.1016/j.aos.2011.11.001.
[3] D. Huang, D. Mu, L. Yang, and X. Cai, "CoDetect: Financial fraud detection with anomaly feature detection," *IEEE Access*, vol. 6, pp. 19161–19174, 2018, doi: 10.1109/ACCESS.2018.2816564.
[4] C. Defang and L. Baichi, "SVM model for financial fraud detection," *J. Northeastern Univ., Natural Sci.*, vol. 40, no. 2, pp. 295–299, Feb. 2019, doi: 10.12068/j.issn.1005-3026.2019.02.027.
[5] *Report to the Nations: 2018 Global Study on Occupational Fraud and Abuse*, ACFE, Austin, TX, USA, Jan. 2019.
[6] J. Sun, H. Fujita, P. Chen, and H. Li, "Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble," *Knowl.-Based Syst.*, vol. 120, pp. 4–14, Mar. 2017, doi: 10.1016/j.knosys.2016.12.019.

[7] R. Saia and S. Carta, "Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks," *Future Gener. Comput. Syst.*, vol. 93, pp. 18–32, Apr. 2019, doi: 10.1016/j.future.2018.10.016.

[8] N. V. Feruleva and M. A. Shtefan, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *J. Corporate Finance Res.*, vol. 11, no. 2, pp. 32–45, Jun. 2017, doi: 10.17323/j.jcfr.2073-0438.11.2.2017.32-45.

[9] H. Xing, G. Zhang, and M. Shang, "Deep learning," *Int. J. Semantic Comput.*, vol. 10, no. 3, pp. 417–439, Sep. 2016, doi: 10.1142/S1793351X16500045.

[10] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, May 2016, doi: 10.1109/TPAMI.2016.2572683.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[12] R. Kanapickienė and Ž. Grundienė, "The model of fraud detection in financial statements by means of financial ratios," *Proc.-Social Behav. Sci.*, vol. 213, pp. 321–327, Dec. 2015, doi: 10.1016/j.sbspro.2015.11.545.

[13] D. Azimi and Y. B. Nahandi, "Investigating the relationship between credit risk indicators and timely fulfillment of the customers' obligations (a case study of bank sepah branches in east Azerbaijan and Ardabil regions)," *Int. J. Accounting Econ. Stud.*, vol. 5, no. 2, pp. 80–86, 2017. [Online]. Available: https://www.sciencepubco.com/index.php/IJAES/article/view/7737, doi: 10.14419/ijaes.v5i2.7737.

[14] X. Zhang, Y. Han, W. Xu, and Q. Wang, "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture," *Inf. Sci.*, vol. 557, pp. 302–316, May 2021, doi: 10.1016/j.ins.2019.05.023.

[15] Y.-J.-J. Goo, D.-J. Chi, and Z.-D. Shen, "Improving the prediction of going concern of Taiwanese listed companies using a hybrid of LASSO with data mining techniques," *SpringerPlus*, vol. 5, no. 1, pp. 539–557, Apr. 2016, doi: 10.1186/s40064-016-2186-5.

[16] R. Maranzato, A. Pereira, M. Neubert, and A. P. do Lago, "Fraud detection in reputation systems in e-markets using logistic regression and stepwise optimization," *ACM SIGAPP Appl. Comput. Rev.*, vol. 11, no. 1, pp. 14–26, Jun. 2010, doi: 10.1145/1869687.1869689.

[17] R. Laptes, A. F. Popa, and F. Dobre, "Research on the evolution of financial audit in Romania-past, current and future trends," *Proc. Econ. Finance*, vol. 15, no. 14, pp. 807–814, Dec. 2014, doi: 10.1016/S2212-5671(14)00523-1.

[18] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 995–1003, May 2007, doi: 10.1016/j.eswa.2006.02.016.

[19] C.-C. Lin, D.-J. Deng, C.-H. Kuo, and L. Chen, "Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers," *IEEE Access*, vol. 7, pp. 56198–56207, 2019, doi: 10.1109/ACCESS.2019.2912631.

[20] S. Tsang, Y. S. Koh, G. Dobbie, and S. Alam, "SPAN: Finding collaborative frauds in online auctions," *Knowl.-Based Syst.*, vol. 71, pp. 389–408, Nov. 2014, doi: 10.1016/j.knosys.2014.08.016.

[21] A. A. Rizki, I. Surjandari, and R. A. Wayasti, "Data mining application to detect financial fraud in Indonesia's public companies," in *Proc. 3rd Int. Conf. Sci. Inf. Technol. (ICSITech)*, Bandung, Indonesia, Oct. 2017, pp. 206–211, doi: 10.1109/ICSITech.2017.8257111.

[22] M. Ausloos, R. Cerqueti, and T. A. Mir, "Data science for assessing possible tax income manipulation: The case of Italy," *Chaos, Solitons Fractals*, vol. 104, pp. 238–256, Nov. 2017, doi: 10.1016/j.chaos.2017.08.012.

[23] K. Pozdniakov, E. Alonso, V. Stankovic, K. Tam, and K. Jones, "Smart security audit: Reinforcement learning with a deep neural network approximator," in *Proc. Int. Conf. Cyber Situational Awareness, Data Analytics Assessment (CyberSA)*, Dublin, Ireland, Jul. 2020, pp. 1–8, doi: 10.1109/CyberSA49311.2020.9139683.

[24] D. N. Rahmatika *et al.*, "Detection of fraudulent financial statement; can perspective of fraud diamond theory be applied to property, real estate, and building construction companies in Indonesia," *Eur. J. Bus. Manage. Res.*, vol. 4, no. 6, pp. 1–12, Nov. 2019, doi: 10.24018/ejbmr.2019.4.6.139.

[25] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," *Decision Support Syst.*, vol. 50, no. 2, pp. 491–500, Jan. 2011, doi: 10.1016/j.dss.2010.11.006.

[26] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud—A comparative study of machine learning methods," *Knowl.-Based Syst.*, vol. 128, pp. 139–152, Jul. 2017, doi: 10.1016/j.knosys.2017.05.001.

[27] C.-C. Lin, A.-A. Chiu, S. Y. Huang, and D. C. Yen, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," *Knowl.-Based Syst.*, vol. 89, pp. 459–470, Nov. 2015, doi: 10.1016/j.knosys.2015.08.011.

[28] K. K. Tangod and G. H. Kulkarni, "Detection of financial statement fraud using data mining technique and performance analysis," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 549–555, Jul. 2015, doi: 10.17148/IJARCCE.2015.47124.

[29] J. Perols, "Financial statement fraud detection: An analysis of statistical and machine learning algorithms," *AUDITING, J. Pract. Theory*, vol. 30, no. 2, pp. 19–50, May 2011, doi: 10.2308/ajpt-50009.

[30] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," *Decision Support Syst.*, vol. 50, no. 3, pp. 595–601, Feb. 2011, doi: 10.1016/j.dss.2010.08.010.

[31] W. Dong, S. Liao, and L. Liang, "Financial statement fraud detection using text mining: A systemic functional linguistics theory perspective," in *Proc. Int. Conf. Pacific Asia Inf. Syst. (PACIS)*, Chiayi, Taiwan, Jun. 2016, pp. 1–5. [Online]. Available: https://aisel.aisnet.org/pacis2016/188

[32] L. Purda and D. Skillicorn, "Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection," *Contemp. Accounting Res.*, vol. 32, no. 3, pp. 1193–1223, Sep. 2015, doi: 10.1111/1911-3846.12089.

[33] P. M. Dechow, W. Ge, C. R. Larson, and R. G. Sloan, "Predicting material accounting misstatements: Predicting material accounting misstatements," *Contemp. Accounting Res.*, vol. 28, no. 1, pp. 17–82, Jan. 2011, doi: 10.1111/j.1911-3846.2010.01041.x.

[34] J. Yao, J. Zhang, and L. Wang, "A financial statement fraud detection model based on hybrid data mining methods," in *Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD)*, Chengdu, China, May 2018, pp. 57–61.

[35] B. An and Y. Suh, "Identifying financial statement fraud with decision rules obtained from modified random forest," *Data Technol. Appl.*, vol. 54, no. 2, pp. 235–255, May 2020, doi: 10.1108/DTA-11-2019-0208.

[36] C. Chimonaki, S. Papadakis, K. Vergos, and A. Shahgholian, "Identification of financial statement fraud in Greece by using computational intelligence techniques," in *Enterprise Applications, Markets and Services in the Finance Industry*, vol. 345. Cham, Switzerland: Springer, May 2019, pp. 39–51, doi: 10.1007/978-3-030-19037-8_3.

[37] P. Craja, A. Kim, and S. Lessmann, "Deep learning application for fraud detection in financial statements," Int. Res. Training Group 1792, High Dimensional Nonstationary Time Ser., IRTG 1792 Discuss. Papers, Humboldt Univ. Berlin, Berlin, Germany, Tech. Rep. 2020-007. [Online]. Available: https://ideas.repec.org/p/zbw/irtgdp/2020007.html

[38] C. L. Jan, "Detection of financial statement fraud using deep learning for sustainable development of capital markets under information asymmetry," *Sustainability*, vol. 13, no. 17, pp. 9879–9898, doi: 10.3390/su13179879.

[39] Accessed: Dec. 12, 2021. [Online]. Available: https://github.com/fxsjy/jieba

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2013, pp. 1–12.

[41] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, vol. 2, Dec. 2013, pp. 3111–3119.

[42] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. Int. Conf. NAACL-HLT*, Atlanta, GA, USA, Jun. 2013, pp. 746–751.

[43] Z. Wang, L. Ma, and Y. Zhang, "A novel method for document summarization using Word2Vec," in *Proc. IEEE 15th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI*CC)*, Palo Alto, CA, USA, Aug. 2016, pp. 523–529.

[44] T. T. Nguyen, H. Tahir, M. Abdelrazek, and A. Babar, "Deep learning methods for credit card fraud detection," Dec. 2020, *arXiv:2012.03754*.

[45] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

[46] D. T. Mane and U. V. Kulkarni, "Visualizing and understanding customized convolutional neural network for recognition of handwritten Marathi numerals," *Proc. Comput. Sci.*, vol. 132, no. 1, pp. 1123–1137, 2018, doi: 10.1016/j.procs.2018.05.027.
[47] A. Cossu, A. Carta, V. Lomonaco, and D. Bacciu, "Continual learning for recurrent neural networks: An empirical evaluation," *Neural Netw.*, vol. 143, pp. 607–627, Nov. 2021, doi: 10.1016/j.neunet.2021.07.021.
[48] Y. J. Kim, B. Baik, and S. Cho, "Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning," *Expert Syst. Appl.*, vol. 62, pp. 32–43, Nov. 2016, doi: 10.1016/j.eswa.2016.06.016.
[49] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1724–1734.
[50] H. A. Ata and I. H. Seyrek, "The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms," *Suleyman Demirel Univ. J. Fac. Econ. Administ. Sci.*, vol. 14, no. 2, pp. 157–170, Feb. 2009.
[51] M. Albashrawi, "Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015," *J. Data Sci.*, vol. 14, no. 3, pp. 553–569, Jul. 2016, doi: 10.6339/JDS.201607_14(3).0010.
[52] I. Dutta, S. Dutta, and B. Raahemi, "Detecting financial restatements using data mining techniques," *Expert Syst. Appl.*, vol. 90, no. 3, pp. 374–393, Dec. 2017, doi: 10.1016/j.eswa.2017.08.030.
[53] C. Liu, Y. Chan, S. H. A. Kazmi, and H. Fu, "Financial fraud detection model: Based on random forest," *Int. J. Econ. Finance*, vol. 7, no. 7, pp. 178–189, Jun. 2015, doi: 10.5539/ijef.v7n7p178.
[54] G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, and R. Ahmad, "A machine learning approach for detection of fraud based on SVM," *Int. J. Sci. Eng. Technol.*, vol. 1, no. 3, pp. 194–198, Jul. 2012.
[55] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization," *J. Inf. Secur. Appl.*, vol. 55, no. 3, pp. 1–15, Dec. 2020, doi: 10.1016/j.jisa.2020.102596.
[56] M. Vasu and V. Ravi, "A hybrid under-sampling approach for mining unbalanced datasets: Applications to banking and insurance," *Int. J. Data Mining Model. Manage.*, vol. 3, no. 1, pp. 75–105, Mar. 2020, doi: 10.1504/IJDMMM.2011.038812.
[57] S. Goel *et al.*, "Can linguistic predictors detect fraudulent financial filings," *J. Emerg. Technol. Accounting*, vol. 7, no. 1, pp. 25–46, Jan. 2010, doi: 10.2308/jeta.2010.7.1.25.
[58] Y. Zhang, M. Sun, Y. Ren, and J. Shen, "Sentiment analysis of sina Weibo users under the impact of super typhoon lekima using natural language processing tools: A multi-tags case study," *Proc. Comput. Sci.*, vol. 174, pp. 478–490, Jan. 2020, doi: 10.1016/j.procs.2020.06.116.
[59] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decis. Support Syst.*, vol. 50, no. 1, pp. 164–175, Dec. 2010, doi: 10.1016/j.dss.2010.07.012.
[60] S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, "Identification of fraudulent financial statements using linguistic credibility analysis," *Decis. Support Syst.*, vol. 50, no. 3, pp. 585–594, Feb. 2011, doi: 10.1016/j.dss.2010.08.009.
[61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.
[62] A. Singh and A. Jain, "Cost-sensitive Metaheuristic technique for credit card fraud detection," *J. Inf. Optim. Sci.*, vol. 41, no. 6, pp. 1319–1331, Sep. 2020, doi: 10.1080/02522667.2020.1809090.
[63] A. Singh and A. Jain, "An empirical study of AML approach for credit card fraud detection-financial transactions," *Int. J. Comput., Commun. Control*, vol. 14, no. 6, pp. 670–690, Nov. 2019, doi: 10.15837/ijccc.2019.6.3498.
[64] C. S. Throckmorton, W. J. Mayew, M. Venkatachalam, and L. M. Collins, "Financial fraud detection using vocal, linguistic and financial cues," *Decis. Support Syst.*, vol. 74, pp. 78–87, Jun. 2015, doi: 10.1016/j.dss.2015.04.006.
[65] S. M. A. Elrahman and A. Abraham, "Class imbalance problem using a hybrid ensemble approach," *Int. J. Hybrid Intell. Syst.*, vol. 12, no. 4, pp. 219–227, Mar. 2016, doi: 10.3233/HIS-160217.
[66] A. C. Turkmen and A. T. Cemgil, "An application of deep learning for trade signal prediction in financial markets," in *Proc. 23nd Signal Process. Commun. Appl. Conf. (SIU)*, May 2015, pp. 2521–2524.
[67] X.-P. Song, Z.-H. Hu, J.-G. Du, and Z.-H. Sheng, "Application of machine learning methods to risk assessment of financial statement fraud: Evidence from China," *J. Forecasting*, vol. 33, no. 8, pp. 611–626, Dec. 2014, doi: 10.1002/for.2294.
[68] X. Liu, "Empirical analysis of financial statement fraud of listed companies based on logistic regression and random forest algorithm," *J. Math.*, vol. 2021, pp. 1–9, Dec. 2021, doi: 10.1155/2021/9241338.
[69] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. 7th Eur. Conf. Knowl. Discovery Databases (PKDD)*, Dubrovnik, Croatia, Sep. 2003, pp. 107–119, doi: 10.1007/978-3-540-39804-2_12.
[70] H. C. Fu, D. J. Chi, and J. Y. Zhu, "Application of random forest, rough set theory, decision tree and neural network to detect financial statement fraud–taking corporate governance into consideration," in *Proc. Int. Conf. Intell. Comput.*, Taiyuan, China, Aug. 2014, pp. 221–234, doi: 10.1007/978-3-319-09333-8_24.
[71] Y. Bao, B. Ke, B. Li, Y. J. Yu, and J. Zhang, "A response to 'critique of an article on machine learning in the detection of accounting fraud,'" *Econ J. Watch*, vol. 18, no. 1, pp. 1–71, Mar. 2021.
[72] Y. Shen, C. Guo, H. Li, J. Chen, Y. Guo, and X. Qiu, "Financial feature embedding with knowledge representation learning for financial statement fraud detection," *Proc. Comput. Sci.*, vol. 187, pp. 420–425, Jan. 2021, doi: 10.1016/j.procs.2021.04.110.

WU XIUGUO was born in Linyi, Shandong, China, in 1975. He received the B.S. degree in computer science and technology from the Shanghai University of Electric Power, Shanghai, in 1999, and the M.S. and Ph.D. degrees in computer application technology from the School of Computer Science and Technology, Shandong University, China, in 2002 and 2010, respectively.

From 2002 to 2010, he was a Lecturer with the School of Information Management, Shandong University of Finance and Economics, China, where he has been an Assistant Professor with the School of Management Science and Engineering, since 2010. His main research interests include business analytics and data mining application in audit, especially text mining based on machine learning.

Dr. Xiuguo was a member of the Chinese Computer Federation. He was a paper reviewer of some journals.

DU SHENGYONG was born in Zaozhuang, Shandong, China, in 1975. He received the B.S. degree in computer science and technology from Shenyang Ligong University, Liaoning, in 1999, and the M.S. degree from the School of Computer Science and Technology, Shandong University, China, in 2006. He is currently pursuing the Ph.D. degree with the Shandong University of Finance and Economics, China.

He is also a Lecturer at the School of Management Science and Engineering, Shandong University of Finance and Economics. His current research interests include swarm intelligence and machine learning.