RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES
Social Science Journal

AI AND MACHINE LEARNING STRATEGIES FOR IDENTIFYING CYBERSECURITY THREATS IN FINANCIAL INSTITUTIONS

S. VENKAT RAO, Assistant Professor
Department of Computer Science and Applications
Sri Ramakrishna Degree and P.G College (A), Nandyal, A.P

ABSTRACT

With the increasing interconnectivity of digital assets, the frequency of cyber threats is rising at an unprecedented pace. Financial institutions should allocate resources towards implementing artificial intelligence-based solutions to detect and safeguard against these threats, thereby safeguarding their assets. Machine learning is an effective tool for analyzing intricate financial security risks that frequently change and can be challenging to forecast. Banks may enhance their comprehension of potential hazards and establish more streamlined data controls by utilizing AI technology like natural language processing, algorithms, and automated reasoning systems. This article proposes the use of artificial intelligence and machine learning to identify cyber security concerns in financial organizations. Machine learning algorithms are continuously enhanced to detect irregularities in the data that could potentially signal a security risk. This methodology empowers financial institutions to detect and protect against malevolent assaults by utilizing tailor-made models that offer practical insights into both internal and external hazards.

1. INTRODUCTION

Financial fraud is the act of using fraudulent and illegal procedures or deceitful strategies to obtain financial benefits. Fraud can occur in various domains of finance, such as banking, insurance, taxation, and business sectors, among others. The prevalence of fiscal fraud and evasion, encompassing activities such as credit card fraud, tax evasion, financial statement fraud, money laundering, and other forms of financial fraud, has been steadily increasing. Despite diligent attempts to eradicate financial fraud, its prevalence has a detrimental impact on both businesses and society, resulting in the loss of hundreds of millions of dollars annually. This substantial financial loss has had a profound impact on individuals, merchants, and banks.

In recent times, there has been a significant surge in fraudulent activities, highlighting the heightened significance of fraud detection. The Association of Certified Fraud Examiners (ACFE) has reported that 10% of white-collar crime cases involve the fraudulent manipulation of financial accounts. The categorization of occupational fraud includes three distinct types: asset misappropriation, corruption, and financial statement fraud. The occurrence of financial statement fraud led to the most substantial losses among them.

While asset misappropriation and corruption occur more frequently than financial statement fraud, the financial consequences of the latter crimes are nevertheless

significantly less severe. According to a survey conducted by Eisner Amper, a prominent accounting firm in the U.S., the average median loss from financial statement fraud in 2018 was $800,000. This amount is more than three times the monetary loss from corruption ($250,000) and seven times the loss from asset misappropriation ($114,000). This study mostly investigates financial statement fraud. Financial statements are comprehensive reports that provide detailed information on a company's commercial activities and financial performance. They include data on income, expenses, profits, loans, potential future issues, and managerial assessments of business success.

All firms are obligated to announce their financial statements in a quarterly and annual manner. Financial statements can be used to indicate the performance of a company. Investors, market analysts, and creditors exploit financial reports to investigate and assess the financial health and earnings potentials of a business. Financial statements consist off our sections; income statement, balance sheet, cash flow statement, and explanatory notes. The income statement places a great emphasis on a company's expenses and revenues during a specific period.

The company's profit or net income is provided in this section, which subtracts expenses from revenues. The balance sheet provides a timely snapshot of liabilities, assets, and stockholders' equity. The cash flow statement measures the extent to which a company is successful in making cash to fund its operating expenses, fund investments, and pay its debt obligations. Explanatory notes are supplemental data that provide clarification and further information about particular items published financial statements of a company.

These notes cover areas including disclosure of subsequent events, asset depreciation, and significant accounting policies, which are necessary disclosures that demonstrate the amounts reported on the financial statements. Financial statement fraud involves falsifying financial statements to pretend the company more profitable than it is, increase the stock prices, avoid payment of the taxes, or get a bank loan.

Fraud triangle in auditing is a framework to demonstrate the motivation behind an individual's decision to commit fraud. The fraud triangle has three elements that increase the risk of fraud: incentive, rationalization, and opportunity, which, together, lead to fraudulent behavior. Auditing professionals have extensively used this theory to explain the motivation behind an individual's decision to commit fraud.

A thorough comprehension of the fraud triangle is essential for assessing instances of financial fraud. Gupta and Singh proposed that the presence of incentives, such as the requirement to accomplish a specific result or compensate for losses, leads to an elevated risk of fraudulent activities. The company may face incentives or pressures to engage in fraudulent actions. Furthermore, the absence of inspections or ineffective controls creates an opportune situation for engaging in fraudulent activities.

Rationalization occurs when a fraudster seeks to provide a justification for their fraudulent behavior, which can be influenced by external factors and circumstances.

2. LITERATURE SURVEY

Evaluation of financial statements fraud detection research: A multidisciplinary analysis

Prior research in the fields of accounting and information systems has shed some light on the significant effects of financial reporting fraud on multiple levels of the economy. In this paper, we compile prior multi-disciplinary literature on financial statement fraud detection. Financial reporting fraud detection efforts and research may be more impactful when the findings of these different domains are combined. We anticipate that this research will be valuable for academics, analysts, regulators, practitioners, and investors.

Interpretable fuzzy rule-based systems for detecting financial statement fraud

Systems for detecting financial statement frauds have attracted considerable interest in computational intelligence research. Diverse classification methods have been employed to perform automatic detection of fraudulent companies. However, previous research has aimed to develop highly accurate detection systems, while neglecting the interpretability of those systems. Here we propose a novel fuzzy rule-based detection system that integrates a feature selection component and rule extraction to achieve a highly interpretable system in terms of rule complexity and granularity. Specifically, we use a genetic feature selection to remove irrelevant attributes and then we perform a comparative analysis of state-of-the-art fuzzy rule-based systems, including FURIA and evolutionary fuzzy rule-based systems. Here, we show that using such systems leads not only to competitive accuracy but also to desirable interpretability. This finding has important implications for auditors and other users of the detection systems of financial statement fraud.

An application of ensemble random forest classifier for detecting financial statement manipulation of Indian listed companies

A rising incidents of financial frauds in recent time has increased the risk of investor and other stakeholders. Hiding of financial losses through fraud or manipulation in reporting and hence resulted into erosion of considerable wealth of their stakeholders. In fact, a number of global companies like WorldCom, Xerox, Enron and number Indian companies such as Satyam, Kingfisher and Deccan Chronicle had committed fraud in financial statement by manipulation. Hence, it is imperative to create an efficient and effective framework for detection of financial fraud. This can be helpful to regulators, investors, governments and auditors as preventive steps in avoiding any possible financial fraud cases. In this context, increasing number of researchers these days have started focusing on developing systems, models and practices to detect fraud in early stage to avoid the any attrition of investor's wealth and to reduces the risk of financing.

In Current study, the researcher has attempted to explore the various 42 modeling

techniques to detect fraud in financial statements (FFS). To perform the experiment, researcher has chosen 86 FFS and 92 non-fraudulent financial statements (nonFFS) of manufacturing firms. The data were taken from Bombay Stock Exchange for the dimension of 2008-2011. Auditor's report is considered for classification of FFS and Non-FFS companies. T-test was applied on 31 important financial ratios and 10 significant variables were taken in to consideration for data mining techniques. 86 FFS and 92 non-FFS during 2008-2017 were taken for testing data set. Researcher has trained the model using data sets. Then, the trained model was applied to the testing data set for the accuracy check. Random forest gives best accuracy. Here, modified random forest model was developed with improved accuracy.

3. EXISTING SYSTEM:

Fraudulent financial statements (FFS) are the results of manipulating financial elements by overvaluing incomes, assets, sales, and profits while underrating expenses, debts, or losses. To identify such fraudulent statements, traditional methods, including manual auditing and inspections, are costly, imprecise, and time-consuming. Intelligent methods can significantly help auditors in analyzing a large number of financial statements. In this study, we systematically review and synthesize the existing literature on intelligent fraud detection in corporate financial statements. In particular, the focus of this review is on exploring machine learning and data mining methods, as well as the various datasets that are studied for detecting financial fraud. We adopted the Kitchen ham methodology as a well-defined protocol to extract, synthesize, and report the results. Accordingly, 47 articles were selected, synthesized, and analyzed. We present the key issues, gaps, and limitations in the area of fraud detection in financial statements and suggest areas for future research. Since supervised algorithms were employed more than unsupervised approaches like clustering, the future research should focus on unsupervised, semi-supervised, as well as bio-inspired and evolutionary heuristic methods for fraud (fraud) detection. In terms of datasets, it is envisaged that future research making use of textual and audio data. While imposing new challenges, this unstructured data deserves further study as it can show interesting results for intelligent fraud detection.

DISADVANTAGES:

- The results is low when compared with proposed.
- Time consumption is high.
- Theoretical limits.

4. PROPOSED SYSTEM

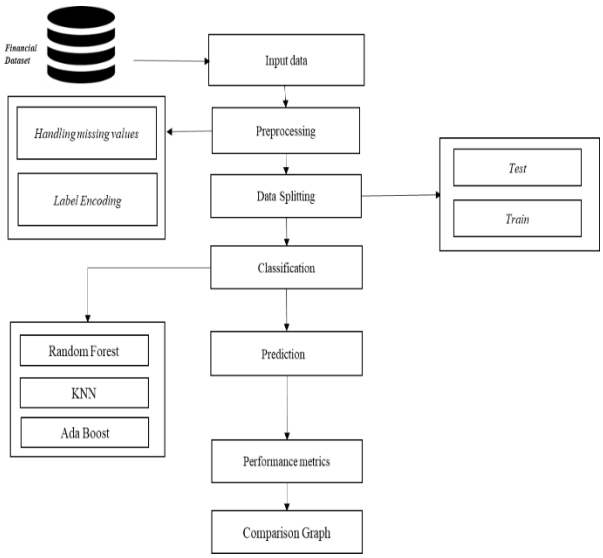
Our suggested solution utilizes a machine learning algorithm to identify fraudulent activity in financial statements. Initially, we choose and examine the imported dataset for future reference. We identify missing values in the dataset and replace them with default values. We are encoding the label within the dataset. We partitioned the dataset into training and testing data in order to predict instances of fraud or non-fraud. Next, we employ three algorithms to enhance accuracy and prediction, ultimately yielding a more precise value. The available algorithms are Random Forest, K-Nearest Neighbors (KNN) classifiers, and AdaBoost Algorithm. Next,

we train the model using the training data from the dataset. Next, we use the training dataset to make predictions on the test dataset. Subsequently, the test values are compared to determine the outcomes of both the actual and predicted values. And we obtain the performance metrics of the dataset. It is imperative to train the models using a dataset that contains both instances of fraud and relevant instances of non-fraud. The ML method is utilized by the system to identify instances as either fraud or non-fraud. The outputs include accuracy, precision, recall, f1-score, and prediction. This demonstrates that the methodology employed in this study can effectively and precisely forecast the likelihood of fraud in the majority of instances. This module provides a straightforward and efficient method to prevent such fraudulent activities and reduce unnecessary expenses.

ADVANTAGES

- It is efficient for large number of datasets.
- The experimental result is high when compared with existing system.
- Time consumption is low.
- Provide accurate prediction results.

5. SYSTEM ARCHITECTURE



6. IMPLEMENTATION

MODULES DESCRIPTION:

DATA SELECTION:

- The input data was collected from the dataset repository like UCI Repository.
- In this process, the input data have some columns like step, type, amount, nameOrig, balanceOrig, nameDest, balanceDest, isFlaggedFraud, etc.

In our collected dataset was read in this process using pandas.

DATA PREPROCESSING:

- Data pre-processing is the process of removing the unwanted data from the dataset.
- Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.
- This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the

accuracy of the dataset, which makes it more efficient.

- Missing data removal
- Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.
- Missing and duplicate values were removed and data was cleaned of any abnormalities.
- Label Encoding: In this process, the string values are converted into integer for more prediction.

Data Splitting

- During the machine learning process, data are needed so that learning can take place.
- In addition to the data required for training, test data are needed to evaluate the performance of the algorithm but here we have training and testing dataset separately.
- In our process, we have to divide as training and testing.
- Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.
- One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.

Classifications

Random Forest Algorithm

- **Random forest** is a machine learning algorithm for fraud detection. It's an unsupervised learning algorithm that identifies fraud by isolating outliers in the data.
- Random Forest is based on the Decision Tree algorithm. It isolates the outliers by randomly selecting a feature from the given set of features and then randomly selecting a split value between the max and min values of that feature.
- This random partitioning of features will produce shorter paths in trees for the fraud data points, thus distinguishing them from the rest of the data.
- Random Forest isolates fraud in the data points instead of profiling non fraud data points. As fraud data points mostly have a lot shorter tree paths than the normal data points, trees in the isolation forest does not need to have a large depth so a smaller max_depth can be used resulting in low memory requirement.

KNN Algorithm:

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a

well suite category by using K- NN algorithm.

- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

7. SCREEN SHOTS

DATA SELECTION:

```
#--------------------Data Selection---------------#
********************
```

	step	type	amount	...	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	NaN	...	0.00	0	0
1	1	PAYMENT	1864.28	...	0.00	0	0
2	1	TRANSFER	NaN	...	0.00	1	0
3	1	CASH_OUT	181.00	...	0.00	1	0
4	1	PAYMENT	11668.14	...	0.00	0	0
5	1	PAYMENT	7817.71	...	0.00	0	0
6	1	PAYMENT	7107.77	...	0.00	0	0
7	1	PAYMENT	7861.64	...	0.00	0	0
8	1	PAYMENT	4024.36	...	0.00	0	0
9	1	DEBIT	5337.77	...	40348.79	0	0
10	1	DEBIT	9644.94	...	157982.12	0	0
11	1	PAYMENT	3099.97	...	0.00	0	0
12	1	PAYMENT	2560.74	...	0.00	0	0
13	1	PAYMENT	11633.76	...	0.00	0	0
14	1	PAYMENT	4098.78	...	0.00	0	0
15	1	CASH_OUT	229133.94	...	51513.44	0	0
16	1	PAYMENT	1563.82	...	0.00	0	0
17	1	PAYMENT	1157.86	...	0.00	0	0
18	1	PAYMENT	671.64	...	0.00	0	0
19	1	TRANSFER	215310.30	...	0.00	0	0

DATA PREPROCESSING

Find Missing Values

```
#--------------------Find missing values---------------#
********************
step              0
type              0
amount            2
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud    0
dtype: int64
```

Handling Missing values:

```
#--------------------Fill 0 from missing Values---------------#
********************
step              0
type              0
amount            0
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud    0
dtype: int64
```

Label Encoding:

```
#--------------------Before Label Encoding---------------#
********************
```

	step	type	amount	...	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	0.00	...	0.00	0	0
1	1	PAYMENT	1864.28	...	0.00	0	0
2	1	TRANSFER	0.00	...	0.00	1	0
3	1	CASH_OUT	181.00	...	0.00	1	0
4	1	PAYMENT	11668.14	...	0.00	0	0
5	1	PAYMENT	7817.71	...	0.00	0	0
6	1	PAYMENT	7107.77	...	0.00	0	0
7	1	PAYMENT	7861.64	...	0.00	0	0
8	1	PAYMENT	4024.36	...	0.00	0	0
9	1	DEBIT	5337.77	...	40348.79	0	0
10	1	DEBIT	9644.94	...	157982.12	0	0
11	1	PAYMENT	3099.97	...	0.00	0	0
12	1	PAYMENT	2560.74	...	0.00	0	0
13	1	PAYMENT	11633.76	...	0.00	0	0
14	1	PAYMENT	4098.78	...	0.00	0	0
15	1	CASH_OUT	229133.94	...	51513.44	0	0
16	1	PAYMENT	1563.82	...	0.00	0	0
17	1	PAYMENT	1157.86	...	0.00	0	0
18	1	PAYMENT	671.64	...	0.00	0	0
19	1	TRANSFER	215310.30	...	0.00	0	0

```
#--------------------After Label Encoding---------------#
********************
```

	step	type	amount	...	newbalanceDest	isFraud	isFlaggedFraud
0	1	3	0.00	...	0.00	0	0
1	1	3	1864.28	...	0.00	0	0
2	1	4	0.00	...	0.00	1	0
3	1	1	181.00	...	0.00	1	0
4	1	3	11668.14	...	0.00	0	0
5	1	3	7817.71	...	0.00	0	0
6	1	3	7107.77	...	0.00	0	0
7	1	3	7861.64	...	0.00	0	0
8	1	3	4024.36	...	0.00	0	0
9	1	2	5337.77	...	40348.79	0	0
10	1	2	9644.94	...	157982.12	0	0
11	1	3	3099.97	...	0.00	0	0
12	1	3	2560.74	...	0.00	0	0
13	1	3	11633.76	...	0.00	0	0
14	1	3	4098.78	...	0.00	0	0
15	1	1	229133.94	...	51513.44	0	0
16	1	3	1563.82	...	0.00	0	0
17	1	3	1157.86	...	0.00	0	0
18	1	3	671.64	...	0.00	0	0
19	1	4	215310.30	...	0.00	0	0

DATA SPLITTING:

```
#--------------------Data Splitting---------------#
*******************
Total no of dataset : (80000, 11)
Training set Without Target (64000, 10)
Training set only Target (64000,)
Testing set Without Target (16000, 10)
Testing set only Target (16000,)
```

CLASSIFICATION:

```
#--------------------Random Forest Algorithm---------------#
*******************
Matrix:
[[15976     0]
 [   12    12]]
classfication:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     15976
           1       1.00      0.50      0.67        24

   micro avg       1.00      1.00      1.00     16000
   macro avg       1.00      0.75      0.83     16000
weighted avg       1.00      1.00      1.00     16000

Accuracy:  99.925
```

```
#--------------------KNN Algorithm---------------#
*******************

Matrix:
[[15975     1]
 [   24     0]]
classfication:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     15976
           1       0.00      0.00      0.00        24

   micro avg       1.00      1.00      1.00     16000
   macro avg       0.50      0.50      0.50     16000
weighted avg       1.00      1.00      1.00     16000

Accuracy:  99.84375
```

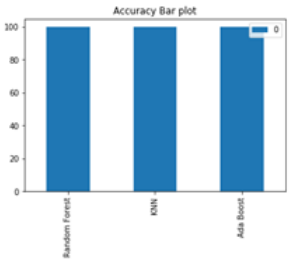
```
#--------------------Ada Boost---------------#
*******************
0.999

Matrix:
[[15973     3]
 [   13    11]]
classfication:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     15976
           1       0.79      0.46      0.58        24

   micro avg       1.00      1.00      1.00     16000
   macro avg       0.89      0.73      0.79     16000
weighted avg       1.00      1.00      1.00     16000

Accuracy:  99.9
```

PREDICTION:

```
#--------------------Get input from user---------------#
*******************

Enter the Step: 1

Enter the Type: 4

Enter the Amount: 0

Enter the nameOrig: 15121

Enter the oldbalance: 181

Enter the newbalance: 0

Enter the nameDest: 7874

Enter the oldbalance: 0

Enter the newbalance: 0

Enter the isFlaggedFraud: 0
[1]
This is financial Fraud
```

GRAPH:

```
#--------------------Camparision batween 3 Algorithm Accuracy---------------#
*******************
```



8. CONCLUSION AND FEATURE ENHANCEMENT

This project presents a method for utilizing the Random Forest algorithm, KNN, and Adaboost algorithm to detect fraud in financial accounts. We refer to the methodology as employing three algorithms on datasets that have been substantially reduced in dimensionality. The Classifications classifier achieves high accuracy results that are equivalent or superior to existing fraud detection algorithms, even when working with limited data and when compared to graph-based methods.

FUTURE ENHANCEMENT

In the future, there will be advancements in the identification of more information through cause-event fraud detection and the

prediction of detection based on cause events. The functionality of the proposed method in a web-based application.

REFERENCES

1. Albizri, D. Appelbaum, and N. Rizzotto, ‘‘Evaluation of financial statements fraud detection research: A multi-disciplinary analysis,’’ Int. J. Discl. Governance, vol. 16, no. 4, pp. 206–241, Dec. 2019.
2. R.Albright, ‘‘Taming text with the SVD.SAS institute white paper, ’’SAS Inst., Cary, NC, USA, White Paper 10.1.1.395.4666, 2004.
3. M. S. Beasley, ‘‘An empirical analysis of the relation between the board of director composition and financial statement fraud,’’ Accounting Rev., vol. 71, pp. 443–465, Oct. 1996.
4. T. B. Bell and J. V. Carcello, ‘‘A decision aid for assessing the likelihood of fraudulent financial reporting,’’ Auditing A, J. Pract. Theory, vol. 19, no. 1, pp. 169–184, Mar. 2000.
5. M.D.BeneishandC.Nichols,‘‘The predictable cost of earnings manipulation,’’Dept.Accounting,Kelley SchoolBus.,IndianaUniv.,Bloomington, IN, USA, Tech. Rep. 1006840, 2007.
6. R. J. Bolton and D. J. Hand, ‘‘Statistical fraud detection: A review,’’ Stat. Sci., vol. 17, no. 3, pp. 235–249, Aug. 2002.
7. 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, 2010.
8. Q. Deng, ‘‘Detection of fraudulent financial statements based on naïve Bayes classifier,’’ in Proc. 5th Int. Conf. Comput. Sci. Educ., 2010, pp. 1032–1035.
9. S. Chen, Y.-J.-J. Goo, and Z.-D. Shen, ‘‘A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements,’’ Sci. World J., vol. 2014, pp. 1–9, Aug. 2014.
10. X. Chen and R. Ye, ‘‘Identification model of logistic regression analysis on listed Firms’ frauds in China,’’ in Proc. 2nd Int. Workshop Knowl. Discovery Data Mining, Jan. 2009, pp. 385–388.
11. Chimonaki, S. Papadakis, K. Vergos, and A. Shahgholian, ‘‘Identification of financial statement fraud in greece by using computational intelligencetechniques,’’inProc.Int.Work shopEnterpriseAppl.,Markets Services Finance Ind. Cham, Switzerland: Springer, 2018, pp. 39–51.
12. R. Cressey, ‘‘Other people’s money; a study of the social psychology of embezzlement,’’ Amer. J. Sociol., vol. 59, no. 6, May 1954, doi: 10.1086/221475.
13. B. Dbouk and I. Zaarour, ‘‘Towards a machine learning approach for earningsmanipulationdetection,’’AsianJ. Bus.Accounting,vol.10,no.2, pp. 215–251, 2017.
14. Q. Deng, ‘‘Application of support vector machine in the detection of fraudulent financial statements,

''inProc.4thInt.Conf.Comput.Sci.Educ., Jul. 2009, pp. 1056–1059.

15. S. Chen, ''Detection of fraudulent financial statements using the hybrid data mining approach,'' SpringerPlus, vol. 5, no. 1, p. 89.