IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 2, 2024

ARTIFICIAL INTELLIGENCE BASED CYBER SECURITY THREATS IDENTIFICATION IN FINANCIAL INSTITUTIONS USING MACHINE LEARNING APPROACH

[1]Chittari Deepika, [2]Mrs. M. Sivamma, [3]Dr.D.William Albert
[1]M.Tech Student, [2]Assistant Professor, [3]Professor
Dept of CSE
Bheema Institute of Technology and Science, Adoni

ABSTRACT

As digital assets become more interconnected, cyber threats are growing at an unprecedented rate. Financial institutions need to invest in artificial intelligence-based solutions for identifying these threats and protecting their assets. Machine learning is a powerful tool for investigating complex financial security threats that constantly evolve and can be difficult to predict. By leveraging AI technologies such as natural language processing, algorithms, and automated reasoning systems, banks can develop a better understanding of potential risks and create more efficient controls around their data. In this paper, an artificial intelligence based cyber security threats identification has proposed in financial institutions using machine learning approach. Machine learning algorithms are constantly being improved to identify anomalies in the data that might indicate a security threat. This approach enables financial firms to identify and defend against malicious attacks using custom-made models that provide actionable insights into both internal and external risks.

1. INTRODUCTION

Financial fraud refers to the use of fraudulent and illegal methods or deceptive tactics to gain financial benefits. Fraud can be committed in different areas of finance, including banking, insurance, taxation, and corporates, and more. Fiscal fraud and evasion, including credit card fraud, tax evasion, financial statement fraud, money laundry, and other types of financial fraud, has become a growing problem. Despite efforts to eliminate financial fraud, its occurrence adversely affects business and society as hundreds of millions of dollars are lost to fraud each year. This significant financial loss has dramatically affected individuals, merchants, and banks.

Nowadays, fraud attempts have increased drastically, which makes fraud detection more important than ever. The Association of Certified Fraud Examiners (ACFE) has announced that 10% of incidents concerning white-collar crime involves falsification of financial statements. They classified occupational fraud into three types: asset misappropriation, corruption, and financial statement fraud. Financial statement fraud resulted in the most significant losses among them.

Although the occurrence frequency of asset misappropriation and corruption is much higher than financial statement fraud, the financial implications of these latter crimes are still far less severe. In particular, as reported in a survey from Eisner Amper, which is among the prominent accounting firms in the U.S., ‘‘the average median loss of financial statement fraud ($800,000 in 2018) accounts for over three times the monetary loss of corruption ($250,000) and seven times as much as asset misappropriation ($114,000)’’.

The focus of this study is on financial statement fraud. Financial statements are documents that describe details about a company, specifically their business activities and financial performance, including income, expenses, profits, loans, presumable concerns that may emerge later, and managerial comments on the business performance.

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.

It is indispensable to understand the fraud triangle to evaluate financial fraud. Gupta and Singh suggested that when there are incentives such as the obligation to achieve an outcome or cover losses, the potential for fraud increases. The company will encounter temptations or pressures to adopt fraudulent practices.

Moreover, the lack of inspections or unsuccessful controls provides a favourable occasion for committing fraud. Rationalization happens when the fraudster aims to justify the fraudulent action, and it could be affected by the others and the conditions.

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.

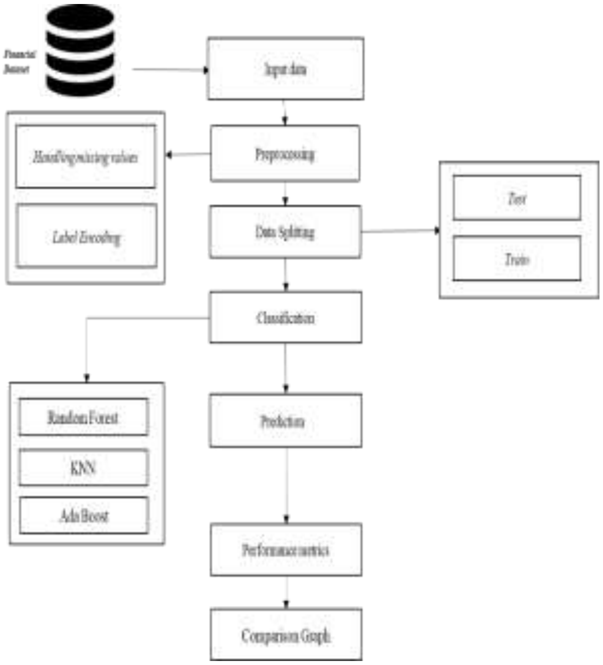
4. PROPOSED SYSTEM

In our proposed system, we detect the fraud in financial statements by using the machine learning algorithm. First, we select and view the imported dataset for future purpose. And we get missing values and fill the default values to the dataset. We encoding the label in the dataset. And we split the dataset to the Train and Test data for predict the fraud or non-fraud. Then we use three algorithms for more accuracy, prediction and which is more accurate value. There are Random forest algorithm, KNN classifiers and Ada-Boost Algorithm. Now, we fit the training data from the dataset. Then we predict the test dataset using training dataset. Then the test values get the results of actual and predicted. And we get the performance of the dataset. It is essential to train the models on data which includes fraud and relevant non fraud. By using the ML algorithm the system is, to classify the fraud and non-fraud and results shows that the accuracy, precision, recall and f1-score and also prediction. This shows that method used in this project can predict the possibility of fraud accurately in most of the cases. This module is the simple and effective way to avoid such frauds and save those expenditures.

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.

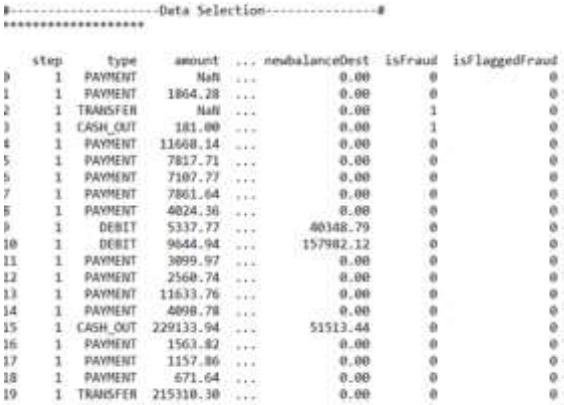
KNN Algorithm:

- 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

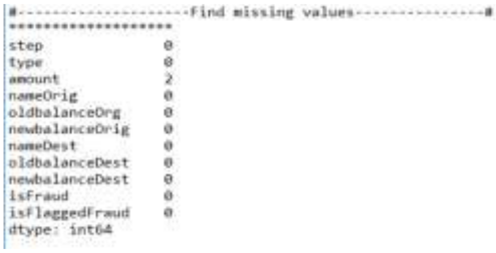
DATA SELECTION:



```
#-------------------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
```

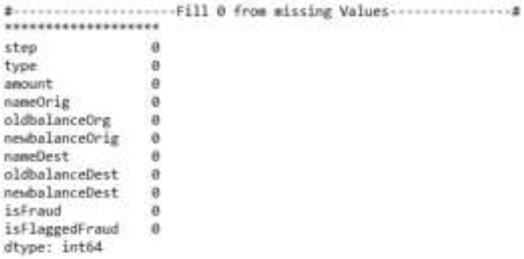
DATA PREPROCESSING

Find Missing Values



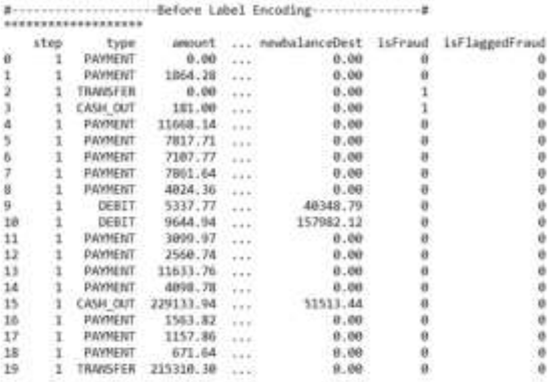
```
#-------------------Find missing values---------------#
*******************
step              0
type              0
amount            2
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud    0
dtype: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

Handling Missing values:

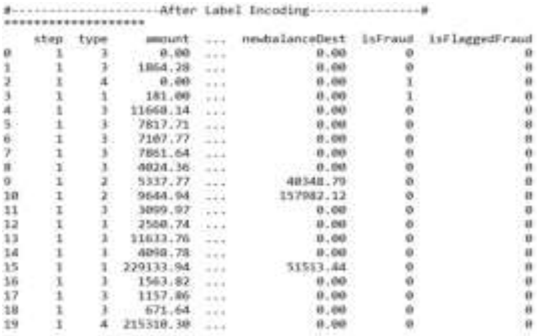


```
#-------------------Fill 0 from missing Values---------------#
*******************
step              0
type              0
amount            0
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud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

Label Encoding:

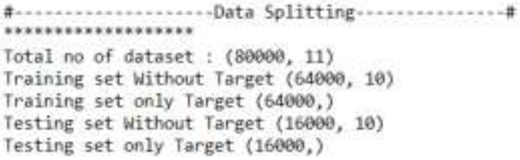


```
#-------------------Before Label Encoding---------------#
*******************
    step      type     amount  ... newbalanceDest  isFraud  isFlaggedFraud
0      1   PAYMENT       0.00  ...           0.00        0               0
1      1   PAYMENT    1864.28  ...           0.00        0               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
```



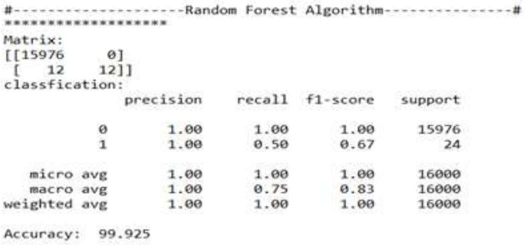
```
#--------------------After Label Incoding---------------#
*******************
    step  type     amount  ...  newbalanceDest  isFraud  isFlaggedFraud
0      1     3       0.00  ...            0.00        0               0
1      1     3    1864.28  ...            0.00        0               0
2      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:

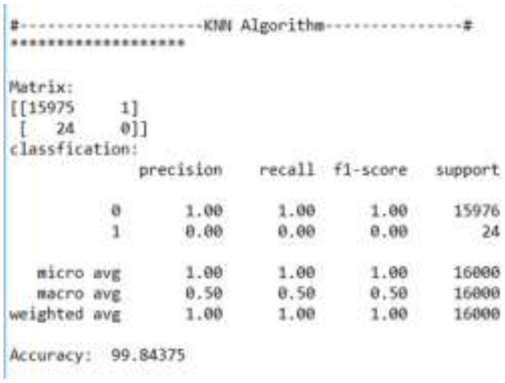


```
#--------------------Data Splitting---------------#
*******************
Total no of dataset : (80000, 11)
Training set Without Target (64000, 10)
Training set only Target (64000,)
Testing set Without Target (16000, 10)
Testing set only Target (16000,)
```

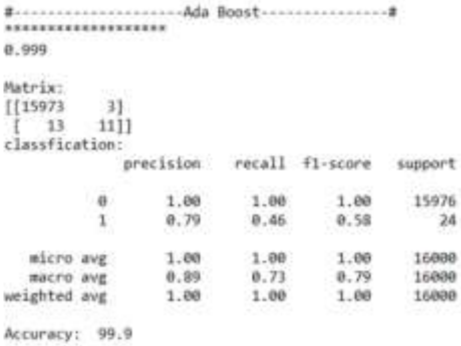
CLASSIFICATION:



```
#--------------------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
```

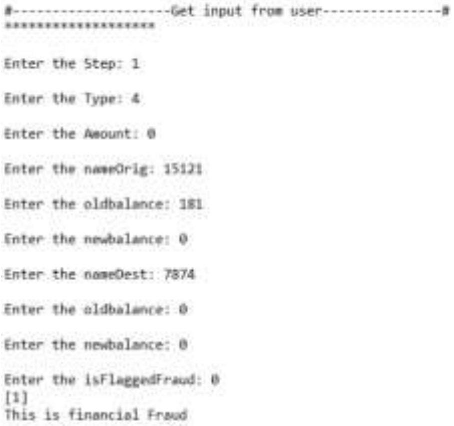


```
#--------------------KNN Algorithm---------------#
*******************

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
```



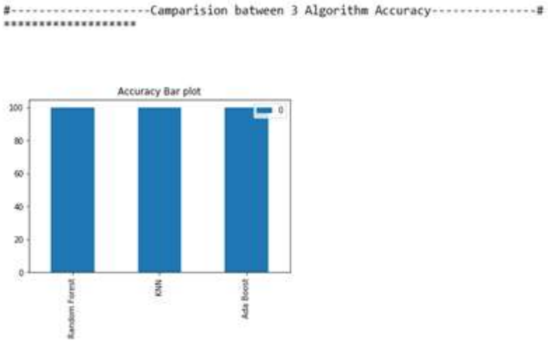
```
#--------------------Ada Boost---------------#
*******************
0.999

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

In this project, we propose an approach to utilise the Random Forest algorithm, KNN and Adaboost algorithm for fraud detection in financial statements. We call the approach the three algorithms on datasets with significantly reduced dimensionality. The Classifications classifier gives high accuracy results that are comparable or superior to other fraud detection techniques in spite of working with reduced data and also compared with graph

FUTURE ENHANCEMENT

In future, discovery of additional information based on cause-event Fraud detection well as prediction of detection based on cause events, etc. The working of the proposed approach in a web 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.