AI-DRIVEN CYBERSECURITY THREAT IDENTIFICATION IN FINANCIAL INSTITUTIONS USING MACHINE LEARNING APPROACHES

[1]T MUTHAIAH, [2]RANGU MANASA
[1]Assistant Professor,[2]Student
Department of CSE
Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

The increasing interconnectedness of digital assets is leading to an unparalleled surge in cyber attacks. Investments in artificial intelligence-based solutions are necessary if financial institutions are to recognise these dangers and safeguard their assets. When examining intricate financial security risks that are dynamic and often unpredictable, machine learning is a potent tool. Through the use of artificial intelligence (AI) technology, such as automated reasoning systems, natural language processing, and algorithms, banks may enhance their comprehension of possible hazards and establish more effective data controls. This study proposes a machine learning technique to identify cyber security concerns in financial institutions using artificial intelligence. Algorithms for machine learning are always being enhanced to find data abnormalities that might point to a security risk. With this strategy, financial institutions may use custom-made models that provide actionable insights into both internal and external threats to detect and fight against harmful assaults.

1. INTRODUCTION

The employment of dishonest, unlawful, or misleading practices to get financial benefits is known as financial fraud. Fraud may occur in a variety of financial contexts, such as banking, insurance, taxes, corporations, and more. A rising issue is fiscal fraud and evasion, which includes money laundering, tax evasion, credit card fraud, financial statement fraud, and other forms of financial fraud. Even with attempts to eradicate financial fraud, hundreds of millions of dollars are lost to it annually, which has a negative impact on society and industry. Banks, retailers, and people have all been severely impacted by this significant financial loss.

These days, there is a marked rise in fraud efforts, which emphasises the need of fraud detection. According to the Association of Certified Fraud Examiners (ACFE), financial statement fraud accounts for 10% of white-collar crime events. They divided occupational fraud into three categories: financial statement fraud, asset theft, and corruption. Among them, financial statement fraud caused the most losses.

Financial statement fraud is significantly less common than asset theft and corruption, but the financial consequences of these crimes are still much less serious. "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)," according to a survey from Eisner Amper, one of the leading accounting firms in the United States.

This research focusses on financial statement fraud. Financial statements provide information on a company's operations and financial performance, including income, costs, profits, loans, potential future problems, and management commentary on the company's performance.

It is mandatory for all companies to provide their financial accounts on a quarterly and yearly basis. A company's performance may be determined by looking at its financial accounts. Financial reports are used by creditors, market analysts, and investors to evaluate a company's profits potential and overall financial health. The income statement, balance sheet, cash flow statement, and explanatory notes make up financial statements. The income statement highlights a company's costs and earnings for a certain time frame.

This section presents the company's profit, or net income, after deducting costs from revenues. An up-to-date picture of the company's assets, liabilities, and shareholders' equity is given by the balance sheet. The cash flow statement evaluates how well a business generates enough cash to cover its debt payments, investments, and operational costs. Explanatory notes are extra details that provide clarification and further details on certain things that are disclosed in a company's financial statements.

The disclosure of later events, asset depreciation, and significant accounting rules are among the topics covered in these notes. These disclosures are essential in order to support the amounts presented on the financial statements. Financial statement fraud is the act of manipulating financial statements to make a firm seem more profitable than it really is, boost stock prices, evade paying taxes, or get a bank loan.

In auditing, the fraud triangle serves as a framework to illustrate the reasons behind a person's choice to commit fraud. The three components of the fraud triangle—opportunity, motivation, and rationalization—all work together to promote fraudulent behaviour and raise the risk of fraud. This hypothesis has been widely used by auditing experts to explain why someone might choose to commit fraud.

To assess financial fraud, knowledge of the fraud triangle is essential. According to Gupta and Singh, the likelihood of fraud rises when there are incentives present, such as the need to meet goals or make up for losses. The business will face pressure or temptation to engage in dishonest business activities.

In addition, the absence of inspections or ineffective controls creates a suitable environment for fraud. When a fraudster attempts to rationalise their fraudulent behaviour, they may take into consideration other people and external factors.

2. LITERATURE SURVEY

Evaluation of financial statements fraud detection research: A multidisciplinary analysis

The substantial repercussions of financial reporting fraud on many levels of the economy have been somewhat illuminated by earlier studies in the disciplines of accounting and information systems. We assemble previous multidisciplinary research on financial statement fraud detection in one study. Combining the results from these several categories may increase the effect of financial reporting fraud detection initiatives and study. We believe that scholars, analysts, regulators, practitioners, and investors will find value in our study.

Interpretable fuzzy rule-based systems for detecting financial statement fraud

Computational intelligence research has seen a significant increase in interest in systems for identifying financial statement fraud. Various categorisation techniques have been used to automatically identify fraudulent businesses. Nevertheless, prior work has focused on creating very precise detection systems, ignoring the systems' interpretability. In order to create a highly interpretable system in terms of rule complexity and granularity, we here offer a unique fuzzy rule-based detection system that combines a feature selection component with rule extraction. To be more precise, we choose features based on genetics to eliminate unnecessary characteristics, and then we conduct a comparison study of the most advanced fuzzy rule-based systems, such as FURIA and evolutionary fuzzy rule-based systems. Here, we demonstrate that the use of such systems yields good interpretability in addition to competitive accuracy. This discovery has significant ramifications for auditors and other users of financial statement fraud detection systems.

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

The danger for investors and other stakeholders has grown recently due to an increase in financial fraud events. concealing financial losses by deception or reporting manipulation, which led to the significant wealth of its stakeholders eroding. As a matter of fact, many multinational corporations such as WorldCom, Xerox, and Enron, as well as some Indian firms including Satyam, Kingfisher, and Deccan Chronicle, were involved in financial statement fraud via manipulation. Therefore, it is essential to establish a framework that is both efficient and successful in detecting financial fraud. Regulators, investors, governments, and auditors may all benefit from this as a preventative measure against potential financial fraud instances. In light of this, a growing number of academics are concentrating on creating methods, models, and systems to identify fraud early on in order to prevent investor wealth loss and lower financing risk.

The researcher's current work aims to investigate 42 different modelling strategies for financial statement (FFS) fraud detection. The researcher selected 86 false financial statements (FFS) and 92 non-FFS financial statements from manufacturing companies to conduct the experiment. The Bombay Stock Exchange provided the statistics for the 2008–2011 timeframe. Consideration is given to the auditor's report for classifying FFS and non-FFS organisations. Thirteen key financial ratios were subjected to the t-test, and data mining methods were used to ten significant variables. For the test data set, 86 FFS and 92 non-FFS were collected between 2008 and 2017. The model has been trained by researchers using datasets. For the purpose of verifying correctness, the trained model was then applied to the testing data set. The most accurate method is random forest. A more accurate version of the random forest model was created in this instance.

3. EXISTING SYSTEM:

The outcome of altering financial aspects is fraudulent financial statements (FFS), which are created by underrating costs, debts, or losses and overvaluing profits, assets, sales, and incomes. Conventional techniques, such as manual audits and inspections, are expensive, inaccurate, and time-consuming for identifying such bogus assertions. When evaluating a large number of financial statements, auditors might benefit greatly from the use of intelligent procedures. We thoroughly examine and summarise the body of research on intelligent fraud detection in company financial statements in this work. This research specifically focusses on investigating machine learning and data mining techniques together with the many datasets that are investigated for financial fraud detection. We used the Kitchen Ham technique as a well-defined procedure to extract, combine, and present the findings. As a result, 47 papers were chosen, combined, and examined. We outline the main problems, weak points, and restrictions in the field of financial statement fraud detection and provide recommendations for further study. Future research on supervised, semi-supervised, bio-inspired, and evolutionary heuristic techniques for fraud detection should be prioritised, since supervised algorithms were used more often than unsupervised approaches such as clustering. It is anticipated that future study will employ both textual and audio information. Even though it presents additional difficulties, further research is necessary since this unstructured data might provide insightful findings for sophisticated fraud detection.

DISADVANTAGES:

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

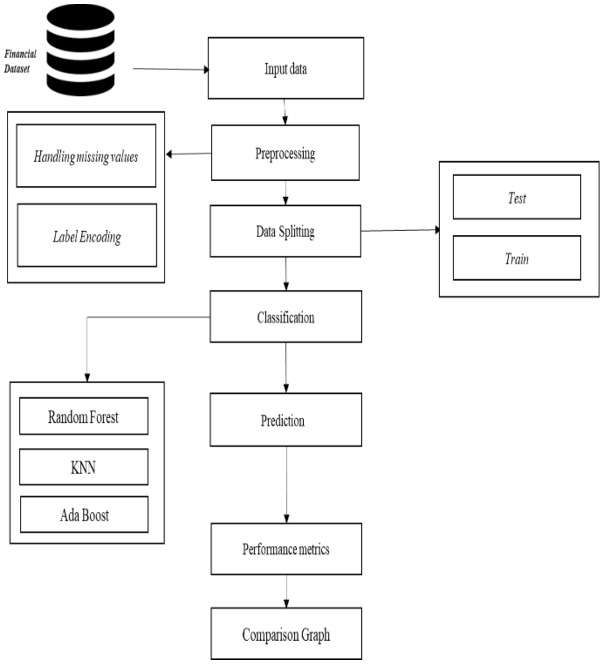
4. PROPOSED SYSTEM

Our suggested approach uses a machine learning algorithm to identify financial statement fraud. To start, we choose and examine the imported dataset for later use. Additionally, we populate the dataset with default values and get missing values. In the dataset, we encoded the label. In order to determine if a dataset is fraudulent or not, we divided it into train and test sets. Then, in order to increase accuracy, prediction, and value accuracy, we use three different methods. There are three algorithms: Ada-Boost, KNN classifiers, and Random Forest. We are now fitting the dataset's training data. Next, we use the training dataset to predict the test dataset. Subsequently, the test values get the expected and actual outcomes. Additionally, we get the dataset's performance. It is essential that the models be trained using both relevant non-fraud data and fraud data. The system classifies fraud and non-fraud using a machine learning algorithm. The accuracy, precision, recall, f1-score, and prediction results are shown. This demonstrates that the approach chosen for this study can, for the most part, forecast the likelihood of fraud with accuracy. This module offers a quick and easy solution to stop these scams and reduce those costs.

ADVANTAGES

- It is low in time consumption, efficient for a large number of datasets, and produces great experimental results when compared to the current system.
- Provide precise forecast outcomes.

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.

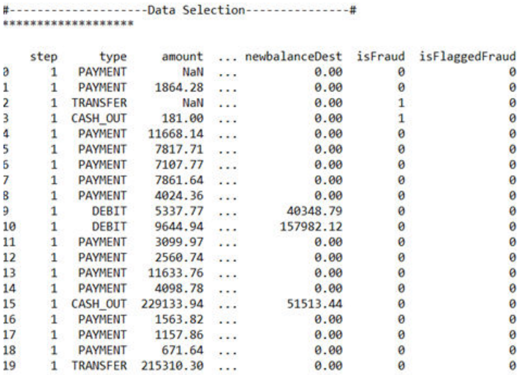
KNN Algorithm:

- 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

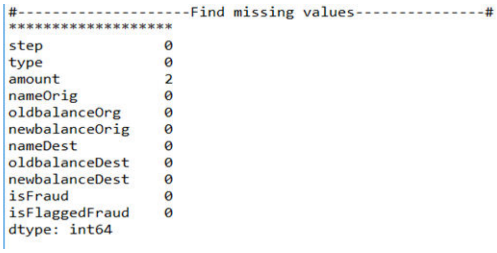
DATA SELECTION:



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

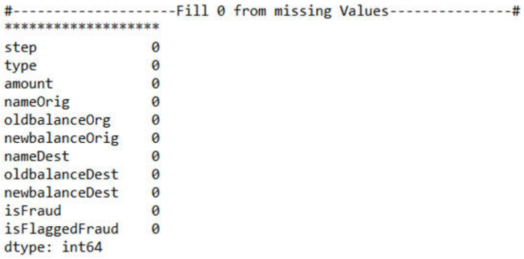
DATA PREPROCESSING

Find Missing Values



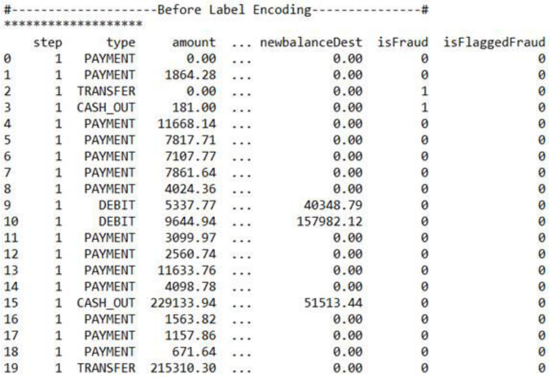
```
#--------------------Find missing values---------------#
******************
step              0
type              0
amount            2
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud    0
dtype: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

Handling Missing values:

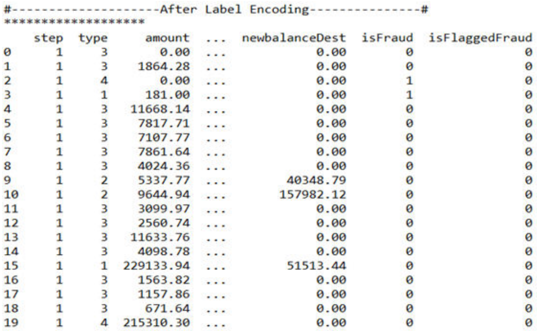


```
#-------------------Fill 0 from missing Values---------------#
******************
step              0
type              0
amount            0
nameOrig          0
oldbalanceOrg     0
newbalanceOrig    0
nameDest          0
oldbalanceDest    0
newbalanceDest    0
isFraud           0
isFlaggedFraud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

Label Encoding:

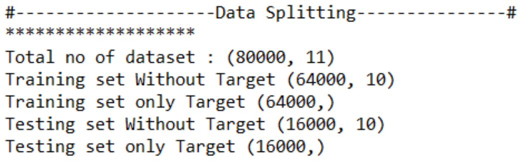


```
#-------------------Before Label Encoding--------------#
******************
    step      type     amount  ... newbalanceDest  isFraud  isFlaggedFraud
0      1   PAYMENT       0.00  ...           0.00        0               0
1      1   PAYMENT    1864.28  ...           0.00        0            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



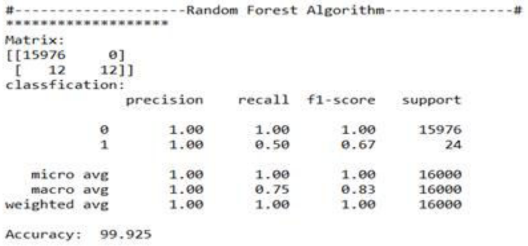
```
#-------------------After Label Encoding--------------#
******************
    step  type     amount  ...  newbalanceDest  isFraud  isFlaggedFraud
0      1     3       0.00  ...            0.00        0               0
1      1     3    1864.28  ...            0.00        0               0
2    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

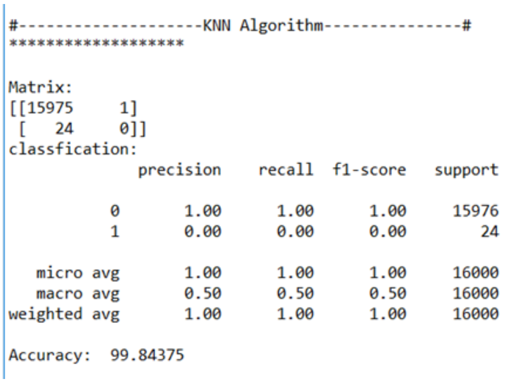


```
#--------------------Data Splitting---------------#
*******************
Total no of dataset : (80000, 11)
Training set Without Target (64000, 10)
Training set only Target (64000,)
Testing set Without Target (16000, 10)
Testing set only Target (16000,)
```

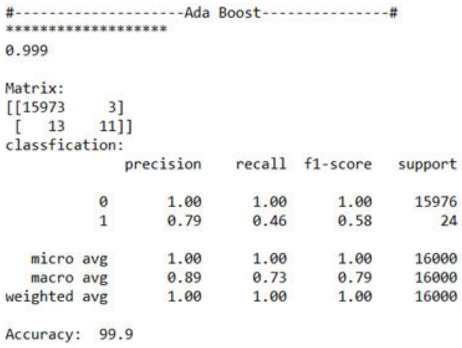
CLASSIFICATION:



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

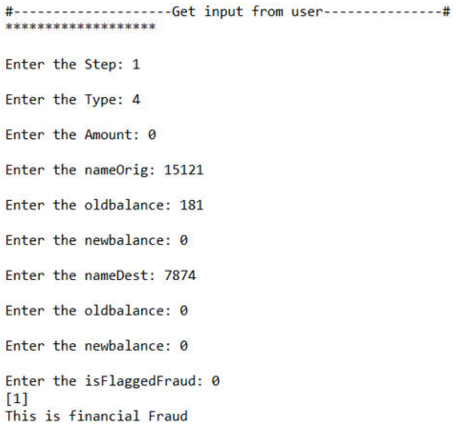


```
#--------------------KNN Algorithm---------------#
*******************

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
```



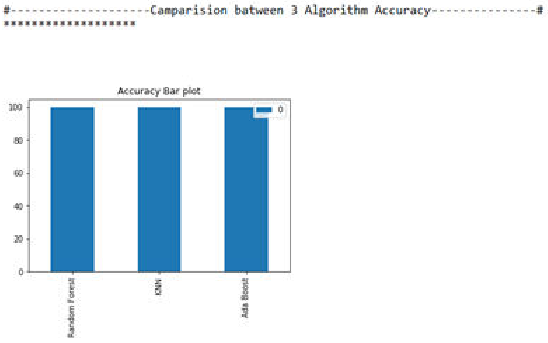
```
#--------------------Ada Boost---------------#
*******************
0.999

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

In this work, we provide a method for detecting fraud in financial accounts by using the Random Forest, KNN, and Adaboost algorithms. We designate the methodology—on datasets with much lower dimensionality—as the three methods. Despite using less data and comparing with graphs, the Classifications classifier produces excellent accuracy results that are on par with or better than other fraud detection algorithms.

FUTURE ENHANCEMENT

In the future, more data may be found via cause-and-event fraud detection as well as detection predictions based on cause events, etc. The suggested method's functionality 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.