

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

¹ P.SINDHURA, ² Mr.A.EMMANUEL RAJU

¹ Student, ² Assistant Professor

Department Of Computer Science & Engineering

Dr. K. V. Subba Reddy Institute Of Technology, Kurnool

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.

I. INTRODUCTION

1.1 General 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

statement places a great emphasis on a company's expenses and revenues during a specific period.

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 of four 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 in 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 is 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.

1.2 Problem Statement:

Fraud detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as fraud, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains.

1.3 Objectives:

The main objective of our project is,

- To predict or to classify the fraud and non-fraud data from financial statements.
- To implement the machine learning algorithm.
- To enhance the performance analysis.

II. LITERATURE SURVEY

2.3.1 Evaluation of financial statements fraud detection research: A multidisciplinary analysis, 2019

Author: A. Albizri, D. Appelbaum, and N. Rizzotto

Methodology

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.

Advantages:

- Reduced Manual power
- Low cost

Disadvantages:

- Too Many False Negatives.
- Run to failure prediction is low.

2.3.2 Interpretable fuzzy rule-based systems for detecting financial statement fraud, 2019**Author: P. Hajek****Methodology**

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.

Advantages:

- Avoid the over fitting from the dataset.

Disadvantages:

- It can be intimidating.

2.3.3 An application of ensemble random forest classifier for detecting financial statement manipulation of Indian listed companies, 2019**Author: H. Patel, S. Parikh, A. Patel, and A. Parikh****Methodology**

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.

Advantages:

- Change of detecting unknown prediction.
- Fraud Detection more efficient than fraud detection, if fraud detection file is large.

Disadvantages:

- Run to failure prediction is low.
- Reliability is unclear.

2.3.4 Detecting fraudulent financial statements for the sustainable development of the socio-economy in China: A multi-analytic approach, 2019

Author: J. Yao, Y. Pan, S. Yang, Y. Chen, and Y. Li

Methodology

Identifying financial statement fraud activities is very important for the sustainable development of a socio-economy, especially in China's emerging capital market. Although many scholars have paid attention to fraud detection in recent years, they have rarely focused on both financial and non-financial predictors by using a multi-analytic approach. The present study detected financial statement fraud activities based on 17 financial and 7 non-financial variables by using six data mining techniques including support vector machine (SVM), classification and regression tree (CART), back propagation neural network (BP-NN), logistic regression (LR), Bayes classifier (Bayes) and K-nearest neighbor (KNN). Specifically, the research period was from 2008 to 2017 and the sample is companies listed on the Shanghai stock exchange and Shenzhen stock exchange, with a total of 536 companies of which 134 companies were allegedly involved in fraud. The stepwise regression and principal component analysis (PCA) were also adopted for reducing variable dimensionality. The experimental results show that the SVM data mining technique has the highest accuracy across all conditions, and after using stepwise regression, 13 significant variables were screened and the classification accuracy of almost all data mining techniques was improved. However, the first 16 principal components transformed by PCA did not yield better classification results. Therefore, the combination of SVM and the stepwise regression dimensionality reduction method was found to be a good model for detecting fraudulent financial statements.

Advantages:

- Rate of missing report is low.
- Simple and Effective method.

Disadvantages:

- Needs to be trained, and trained model carefully otherwise tends to be false positive.
- Low Accuracy rate.

2.3.5 An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning, 2020

Author: Wu Xiuguo, Du Shengyong

Methodology

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.

Advantages:

- Flexibility, fault tolerance, high sensing fidelity, low-cost and rapid deployment.

Disadvantages:

- Sensor nodes are prone to failures.

2.3.5 Fraud Detection in Financial Statements using Text Mining Methods: A Review, 2021

Autor: Ajit Kr. Singh Yadav and Marpe Sora Methodology

In the financial industry, financial fraud is one of the ever-growing hazards with far concerns. Financial statements are the fundamental papers which replicate economic position of a corporation. Users of the financial information like public, creditors etc. are the major foundations of a decision-making process for financing stakeholders. Financial fraud has extremely damaged the sustainable growth of financial markets and enterprises. The amount of financial reporting fraud cases keeps on developing. Each incident is a thick hit to partners, banks, and financial specialists and it costs human progress significantly. One of the serious issues is to recognize the financial reporting fraud by utilizing formation of an active model. The aim of this paper is to identifying frauds using various text mining techniques and guard, the public's investments. This investigation will benefit auditors and financial governors.

Advantages:

- Simplest and Easiest Data Mining Approach.

Disadvantages:

- Handling of Fraud detection is difficult

III. SYSTEM ANALYSIS

3.1 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.

3.1.1 DISADVANTAGES:

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

3.2. 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.

3.2.1 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.

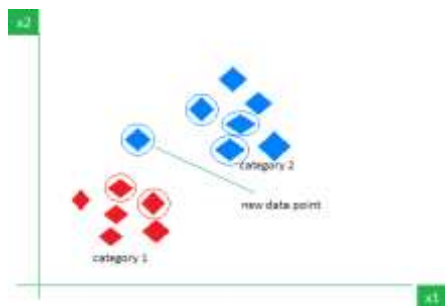
IV. ALGORITHMS USED AND BUILDING MODEL

K-Nearest Neighbors Algorithm?

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

As an example, consider the following table of data points containing two features:

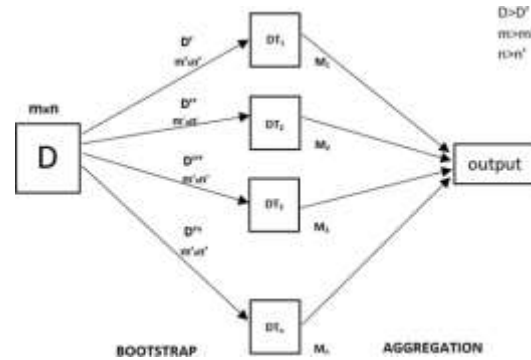


Now, given another set of data points (also called testing data), allocate these points to a group by analyzing the training set. Note that the unclassified points are marked as 'White'.

Random Forest

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision

tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation.



Random Forest Regression?

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

We need to approach the Random Forest regression technique like any other machine learning technique.

- Design a specific question or data and get the source to determine the required data.
- Make sure the data is in an accessible format else convert it to the required format.
- Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
- Create a machine-learning model.

- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn't satisfy your expectations, you can try improving your model accordingly or dating your data, or using another data modeling technique.
- At this stage, you interpret the data you have gained and report accordingly.

ADABOOST

Adaboost is one of the earliest implementations of the boosting algorithm. It forms the base of other boosting algorithms, like gradient boosting and xgboost.

This tutorial will take you through the math behind implementing this algorithm and also a practical example of using the scikit-learn adaboost api.

What is boosting

Boosting is a general ensemble method that creates a strong classifier from a number of weak ones. We first build a weak model and then build a second model based on the errors from the first model. This process is repeated over and over again until we build a classifier that can make predictions accurately and the error is minimised.

Boosting differs from bagging in that it trains the weak learners sequentially and not in parallel. This process can be described as:

- Train the model h1 on the whole set
- Train the model h2 with exaggerated data on the regions in which h1 performs poorly
- Train the model h3 with exaggerated data on the regions in which $h1 \neq h2$... and so on

AdaBoost

AdaBoost is one of the first boosting algorithms to have been introduced. It is mainly used for classification, and the base learner (the machine learning algorithm that is boosted) is usually a decision tree with only one level, also called as stumps.

It makes use of weighted errors to build a strong classifier from a series of weak classifiers.

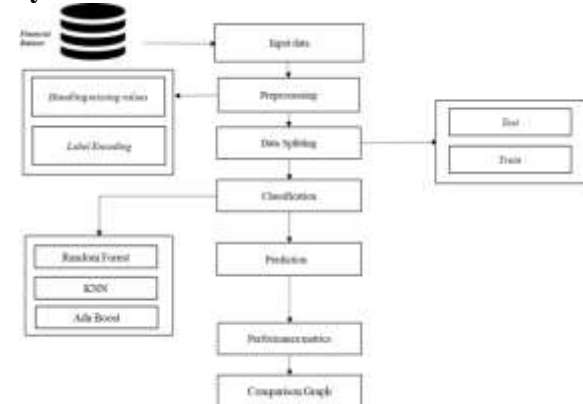
Algorithm behind Adaboost

It works in the following steps:

1. Initially, Adaboost selects a training subset randomly
2. It iteratively trains the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training
3. It assigns the higher weight to wrong classified observations so that in the next iteration these observations will get the high probability for classification
4. Also, It assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight
5. This process iterates until the complete training data fits without any error or until reached to the specified maximum number of estimators

V. SYSTEM DESIGN

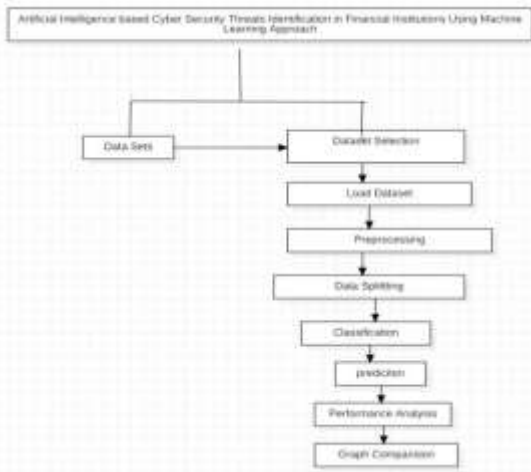
System Architecture



The proposed botnet detection model based on machine learning using DNS query data. The model is built on the analysis that Threats of CS Threats routinely send lookup queries to the DNS system to find IP addresses of C & C servers using automatically generated domain names. The detection model is implemented in two phases: (a) the training phase and (b) the detection phase. During the training phase, the DNS query data is collected, and then domain names in DNS queries are extracted. Next, the set of domain names is pre-processed to extract the features for the training. In the training phase, machine learning algorithms are used to learn the classifiers. Through the evaluation

process, the machine learning algorithm that gives the highest overall classification accuracy will be selected for use in the proposed detection model. During the detection phase of the model, the DNS queries are monitored and passed through the process of extracting the domain names, pre-processing, and classifying using the classifier produced from the training phase to determine if a domain name is legitimate, or a Threats domain name. The pre-processing step for each domain name in the training and detection phase is the same. However, this step is done in the offline mode for all the domain names of the training dataset in the training phase while it is done for each domain name extracted from the DNS query on the fly in the detection phase.

VI. DETAILED DESCRIPTION BLOCK DIGRAM



VII. IMPLEMENTATION MODULES DESCRIPTION

- Data selection
- Data preprocessing
- Data splitting
- Classification
- Prediction
- Performance Metrics
- Graph Comparison

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.

Ada Boost Algorithm:

- AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps.
- AdaBoost is best used to boost the performance of decision trees on binary classification problems.
- AdaBoost was originally called AdaBoost.M1 by the authors of the technique Freund and Schapire. More recently it may be referred to as discrete AdaBoost because it is used for classification rather than regression.
- AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem.
- The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level. Because these trees are so short and only contain one decision for classification, they are often called decision stumps.

Prediction

- Predict the dataset values are Non Fraud/Fraud by using classification algorithm.

Performance Metrics

The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

- **Accuracy:** Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

$$AC = (TP+TN) / (TP+TN+FP+FN)$$

- **Precision:** Precision is defined as the number of true positives divided by the number

of true positives plus the number of false positives.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

- **Recall:** Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

- **F1-score:** F1 score of the positive class in binary classification or weighted average of the F1 scores of each class for the multiclass task. When true positive + false positive == 0, precision is undefined. When true positive + false negative == 0, recall is undefined.

$$\text{F1-score} = \frac{2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))}{1}$$

Graph Comparison:

Comparison between 3 algorithms for

VIII. 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    7187.77     0.00      0          0
7      1  PAYMENT    7861.64     0.00      0          0
8      1  PAYMENT    4824.36     0.00      0          0
9      1  DEBIT      5337.77     48348.79   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
oldbalanceOrig 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
oldbalanceOrig 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    7187.77     0.00      0          0
7      1  PAYMENT    7861.64     0.00      0          0
8      1  PAYMENT    4824.36     0.00      0          0
9      1  DEBIT      5337.77     48348.79   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          7187.77     0.00      0          0
7      1  3          7861.64     0.00      0          0
8      1  3          4824.36     0.00      0          0
9      1  2          5337.77     48348.79   0          0
10     1  2          9644.94     157982.12  0          0
11     1  3          3099.97     0.00      0          0
12     1  1          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,)
```

IX. 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. Beneish and C. Nichols, "The predictable cost of earnings manipulation," *Dept. Accounting, Kelley School Bus., Indiana Univ., 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 intelligence techniques," in *Proc. Int. Workshop Enterprise Appl., 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 earnings manipulation detection," *Asian J. 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," in *Proc. 4th Int. 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, Dec. 2011.