**Stock Market Prediction:-**

**Using Historical Data Analysis**

The stock market prediction process is filled with uncertainty and can be influenced by multiple factors. Social media data has a high impact due to its increased usage, and it can be helpful in predicting the trend of the stock market. Technical analysis is done

using by applying machine learning algorithms on historical data of stock prices. The method usually involves gathering various social media data, news to extract sentiments expressed by individuals. Other data like previous year stock prices are also considered. The relationship between various data points is considered, and a prediction is made on these data points. The model was able to make predictions about future stock values.

In this proposed system, we focus on predicting the stock values using machine learning algorithms like Random Forest and Support Vector Machines. We proposed the system “Stock market price prediction” . We have predicted the stock market price using the random forest algorithm.

**SVM**

Support Vector Machine commonly known as SVM provides the kernel, decision function, and sparsity of the solution. It is used to learn polynomial radial basis function and the multi-layer perceptron classifier. It is a training algorithm for classification and regression, which works on a larger dataset. There are many algorithms in the market but SVM provides better efficiency and accuracy. The correlation analysis between SVM and the stock market indicates strong interconnection between the stock prices and the market index.

Financial organizations and merchants have made different exclusive models to attempt and beat the market for themselves or their customers, yet once in a while has anybody accomplished reliably higher-than-normal degrees of profitability. Nevertheless, the challenge of stock forecasting is so engaging in light of the fact that the improvement of only a couple of rate focuses can build benefit by a large number of dollars for these organizations.

**A Stock Market Prediction Method Based on Support Vector Machines (SVM) and Independent Component Analysis (ICA)**

The time series prediction problem was researched in the work centers in the various financial institutions. The prediction model, which is based on SVM and independent analysis, combined is called SVM-ICA.

The SVM is designed to solve regression problems in nonlinear classification and time series analysis. The generalization error is minimized using an approximate function, which is based on risk diminishing principle. Thus, the ICA technique extracts various important features from the dataset. The time series prediction is based on SVM. The result of the SVM model was compared with the results of the ICA technique without using a preprocessing step.

Results were tried on the binary classification utilizing SVM classifier with an alternate set of a feature list. The greater part of the Machine Learning approach for taking care of business issues had their benefit over factual techniques that did exclude AI, despite the fact that there was an ideal procedure for specific issues. The **Swarm Intelligence optimization method named Cuckoo** **search** was most easy to accommodate the parameters of SVM. The proposed hybrid CS-SVM strategy exhibited the performance to create increasingly exact outcomes in contrast with ANN. Likewise, the CS-SVM display performed better in the forecasting of the stock value prediction. Prediction stock cost utilized to parse records to compute the predicted, send it to the user, and autonomously perform tasks like buying and selling shares utilizing automation concepts. Naive Bayes Algorithm was utilized.

**Random Forest Algorithm**

Random forest algorithm is being used for the stock market prediction. Since it has been termed as one of the easiest to use and flexible machine learning algorithms, it gives good accuracy in the prediction. This is usually used in the classification tasks. Because of the high volatility in the stock market, the task of predicting is quite challenging. In stock market prediction we are using a random forest classifier which has the same hyperparameters as of a decision tree.The decision tool has a model similar to that of a tree. It takes the decision based on possible consequences, which includes variables like event outcome, resource cost, and utility. The random forest algorithm represents an algorithm where it randomly selects different observations and features to build several **decision trees** and then takes the aggregate of the several decision trees outcomes. The data is split into partitions based on the questions on a label or an attribute. The data set we used was from the previous year‟s stock markets collected from the public database available online, 80 % of data was used to train the machine and the rest 20 % to test the data. The basic approach of the supervised learning model is to learn the patterns and relationships in the data from the training set and then reproduce them for the test data.

**Support Vector Machine Algorithm**

The main task of the support machine algorithm is to identify an N-dimensional space that distinguishably categorizes the data points. Here, N stands for a number of features. Between two classes of data points, there can be multiple possible hyperplanes that can be chosen. The objective of this algorithm is to find a plane that has maximum margin. Maximizing margin refers to the distance between data points of both classes. The benefit associated with maximizing the margin is that it provides some reinforcement so that future data points can be more easily classified. Decision boundaries that help classify data points are called hyperplanes. Based on the position of the data points relative to the hyperplane they are attributed to different classes. The dimension of the hyperplane relies on the number of attributes, if the number of attributes is two then the hyperplane is just a line, if the number of attributes is three then the hyperplane is two dimensional.

**METHODOLOGIES**

1. Classification is an instance of supervised learning where a set is analyzed and categorized based on a common attribute. From the values or the data are given, classification draws some conclusion from the observed value. If more than one input is given then classification will try to predict one or more outcomes for the same. A few classifiers that are used here for the stock market prediction includes the random forest classifier, SVM classifier.

**Random Forest Classifier**

Random forest classifier is a type of ensemble classifier and also a supervised algorithm. It basically creates a set of decision trees, that yields some result. The basic approach of random class classifiers is to take the decision aggregate of random subset decision trees and yield a final class or result based on the votes of the random subset of decision trees.

**Parameters**

The parameters included in the random forest classifier are n\_estimators which is total number of

decision trees, and other hyper parameters like oobscore to determine the generalization accuracy of the random forest, max\_features which includes the number of features for best-split. min\_weight\_fraction\_leaf is the minimum weighted fraction of the sum total of weights of all the input samples required to be at a leaf node. Samples have equal weight when sample weight is not provided.

**SVM classifier**

SVM classifier is a type of discriminative classifier.

The SVM uses supervised learning i.e. a labeled training data. The output is hyperplanes which categorizes the new dataset. They are supervised learning models that use associated learning algorithms for classification and as well as regression.

**Parameters**

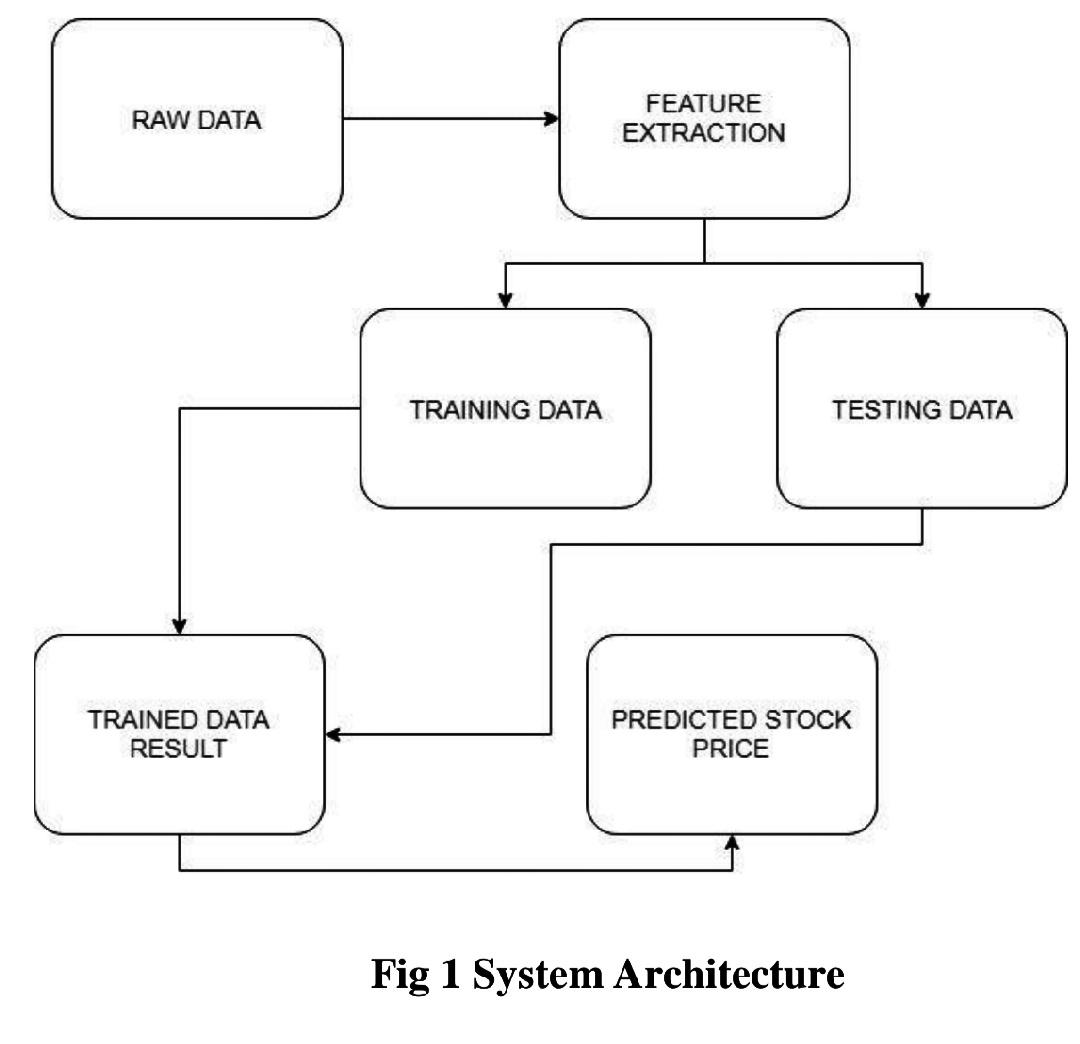
**The tuning parameters of SVM classifiers are kernel parameters, gamma parameters and regularization parameters**. Kernels can be categorized as linear and polynomial kernels calculate the prediction line.

linear kernels prediction for a new input is calculated by the dot product between the input and the support vector.

*C parameter* is known as the regularization parameter; it determines whether the accuracy of the model increases or decreases. The default value of c =10.Lower regularization value leads to misclassification. Gamma parameters measure the influence of a single training on the model. Low values signifies far from the plausible margin and high values signifies closeness from the plausible margin.

**SYSTEM ARCHITECTURE**

Kaggle is an online community for data analysis and predictive modeling. It also contains a dataset of different fields, which is contributed by data miners. Various data scientists compete to create the best models for predicting and depicting the information. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges. The dataset used in the proposed project has been downloaded from Kaggle. However, this data set is present in what we call raw format. The data set is a collection of stock market information about a few companies. The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So the first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction starts from an initial state of measured data and builds derived values or features. These features are intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction is a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for ease of management, while still precisely and totally depicting the first informational collection. The feature extraction process is followed by a classification process wherein the data that was obtained after feature extraction is split into two different and distinct segments. Classification is the issue of recognizing to which set of categories a new observation belongs. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model. The splitting is done in a way that training data maintain a higher proportion than the test data. The random forest algorithm utilizes a collection of random decision trees to analyze the data. In layman terms, from the total number of decision trees in the forest, a cluster of the decision trees look for specific attributes in the data. This is known as data splitting. In this case, since the end goal of our proposed system is to predict the price of the stock by analyzing its historical data.



**MODULE IDENTIFICATION**

The various modules of the project would be divided into the segments as described-:

**I. Data Collection**

Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects. Data collection also complements to enhance the dataset by adding more data that is external. Our data mainly consists of the previous year stock prices. Initially, we will be analyzing the obtained from yahoo! dataset and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

**II. Pre Processing Data**

pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually inconsistent or incomplete and usually contains many errors. The data pre-processing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test sets and finally doing a feature scaling to limit the range of variables so that they can be compared on common environs.

**III. Training the Machine**

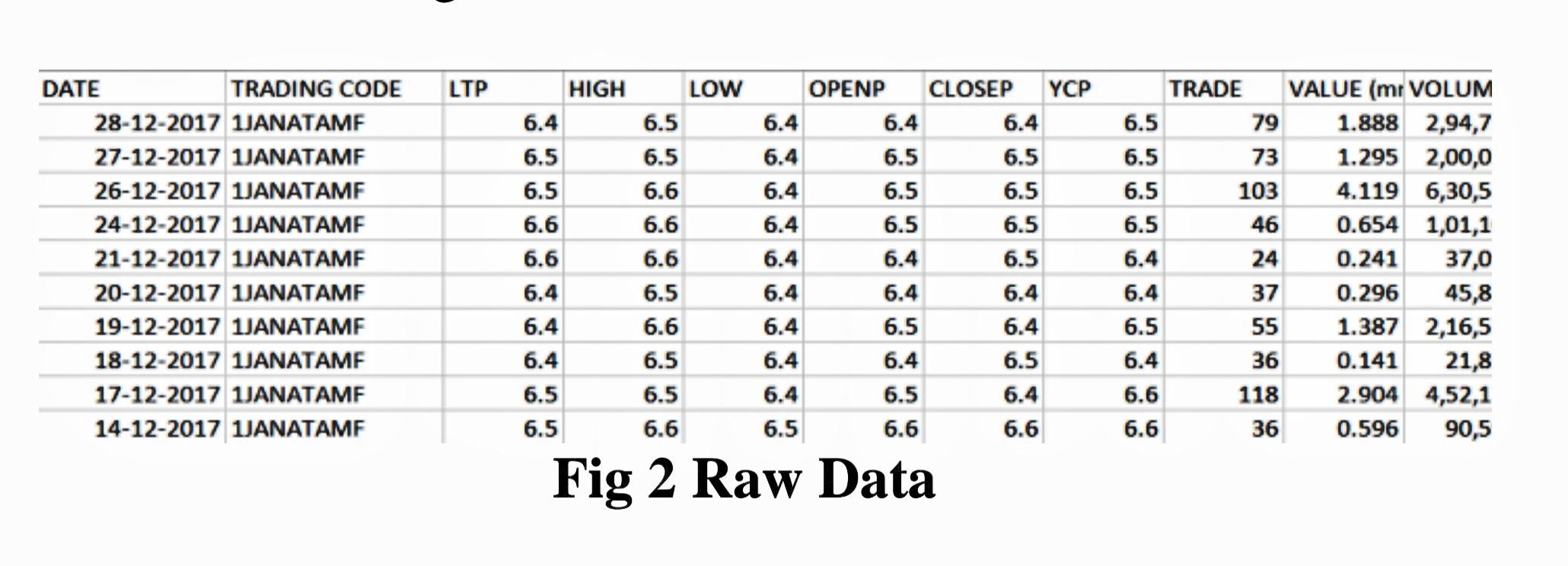
Training the machine is similar to feeding the data to the algorithm to touch up the test data. The Training sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes cross-validation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyperparameters like the number of trees in a random forest. We perform the entire cross-validation loop on each set of hyperparameter values. Finally, we will calculate a cross-validated score, for individual sets of hyperparameters. Then, we select the best hyperparameters. The idea behind the training of the model is that we take some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data.

**IV. Data Scoring**

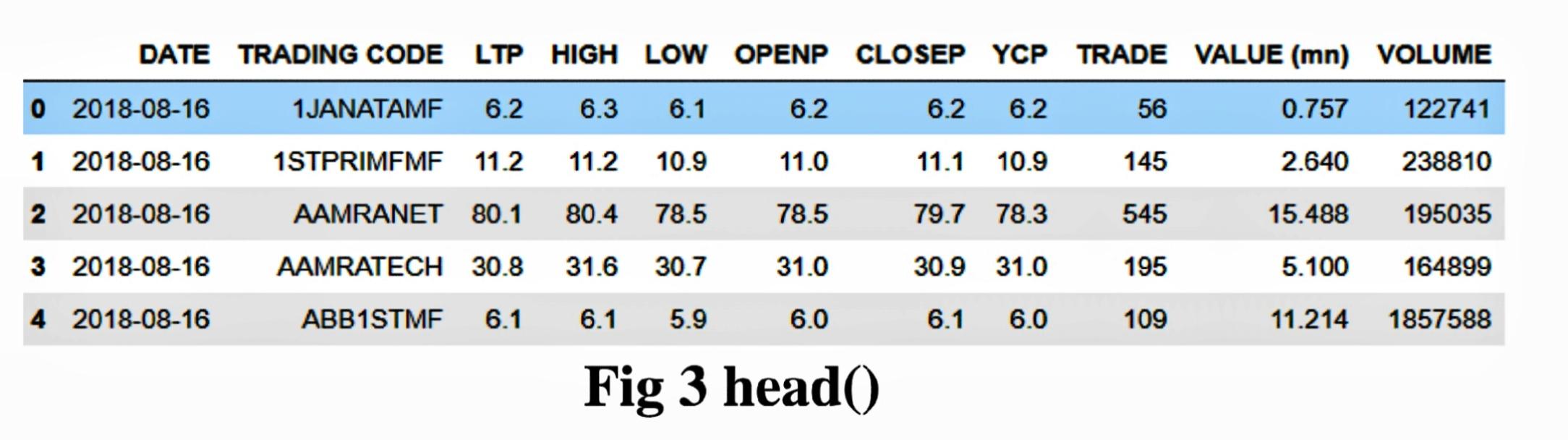
The process of applying a predictive model to a set of data is referred to as scoring the data. The technique used to process the dataset is the Random Forest Algorithm. Random forest involves an ensemble method, which is usually used, for classification and as well as regression. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

**EXPERIMENTAL RESULTS**

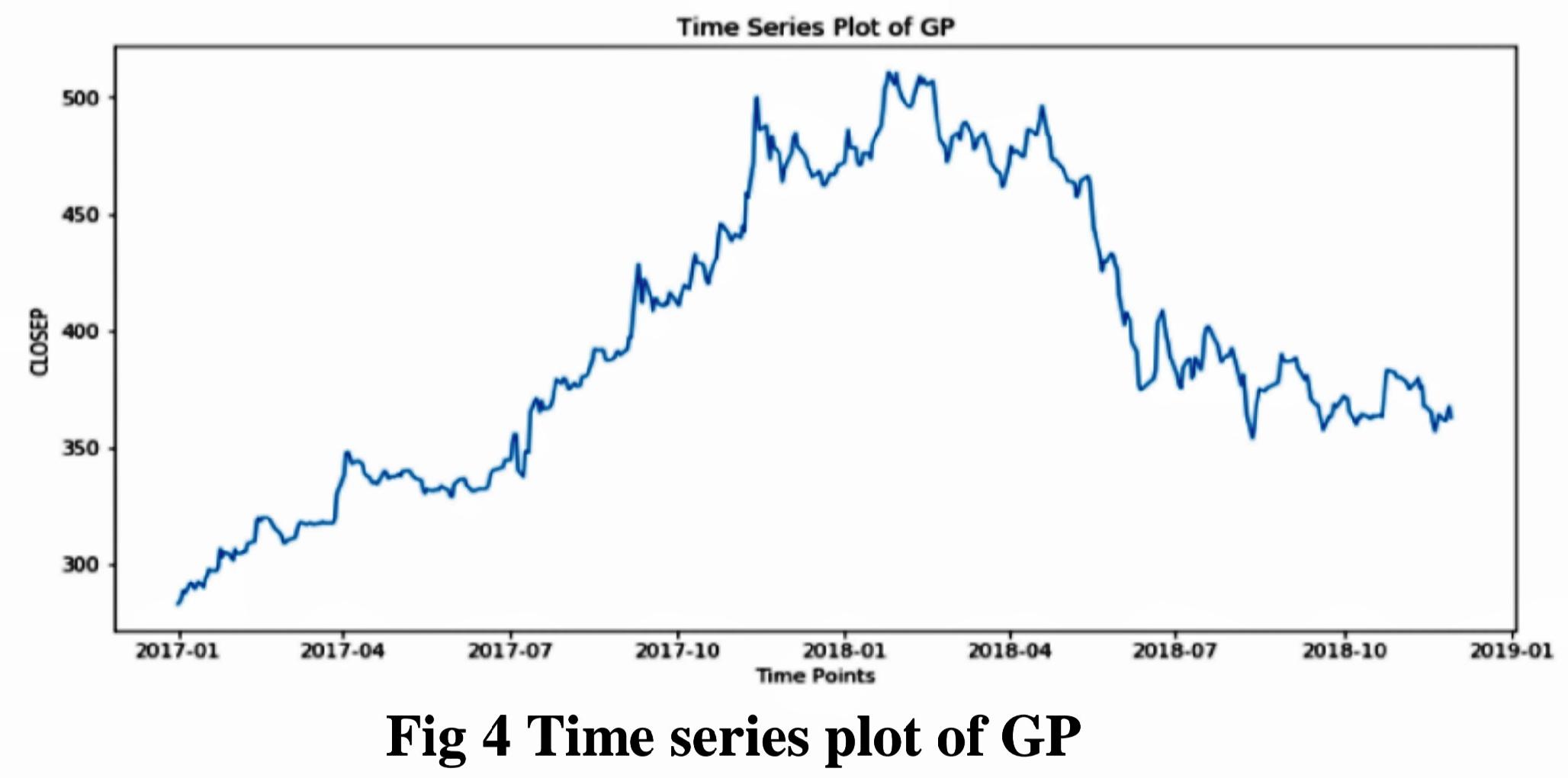
The xlxs file contains the raw data based on which we are going to publish our findings. There are eleven columns or eleven attributes that describe the rise and fall in stock prices. Some of these attributes are (1) HIGH, which describes the highest value the stock had in the previous year. (2) LOW, is quite the contrary to HIGH and resembles the lowest value the stock had in previous year (3) OPENP is the value of the stock at the very beginning of the trading day, and (4) CLOSEP stands for the price at which the stock is valued before the trading day closes. There are other attributes such as YCP, LTP, TRADE, VOLUME and VALUE, but the above mentioned four play a very crucial role in our findings.



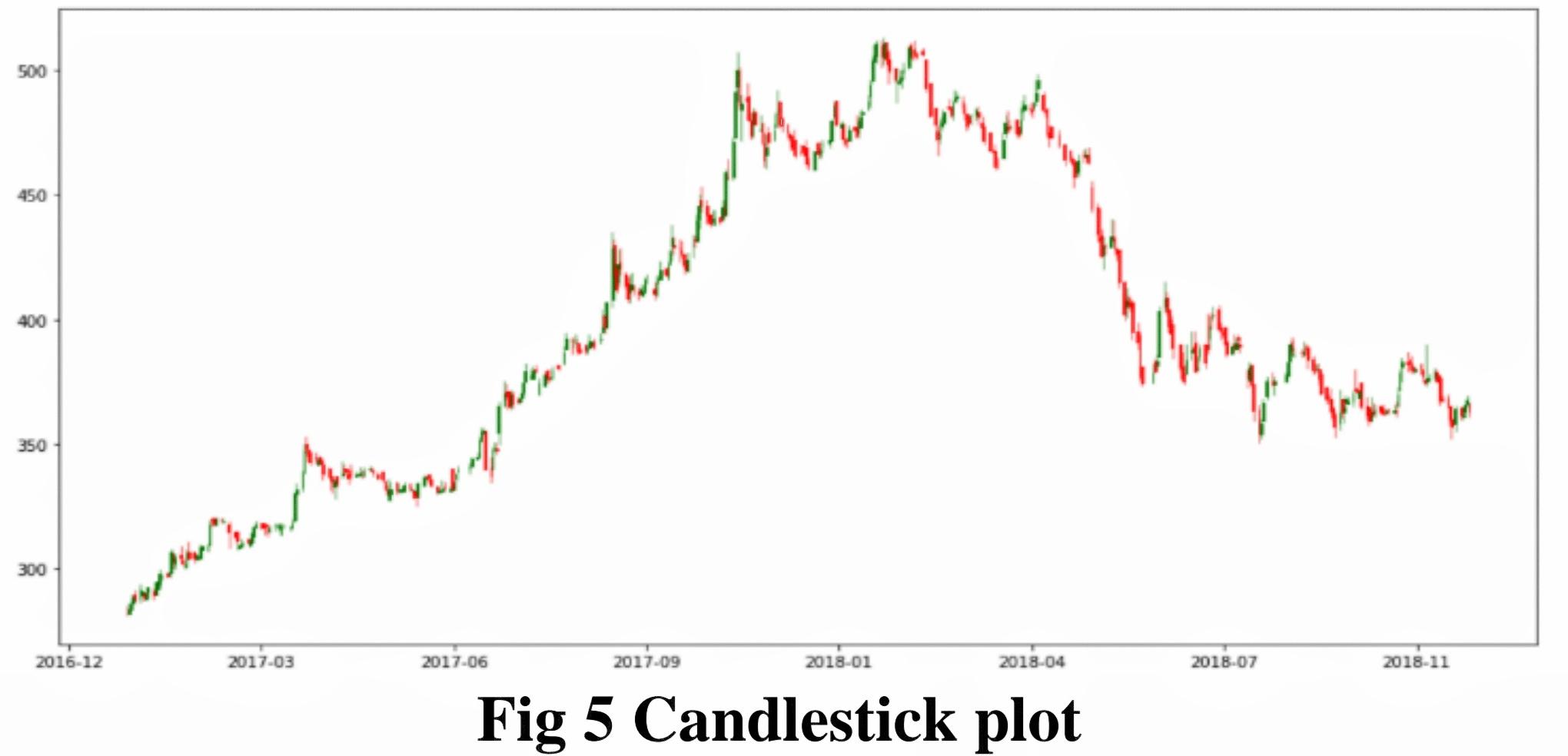
Raw Data This is a pictorial representation of the data present in our xlxs file. This particular file contains 121608 such records. There are more than ten different trading codes available in the dataset and some of the records do not have relevant information that can help us train the machine, so the logical step is to process the raw data. Thus we obtain a more refined dataset which can now be used to train the machine.

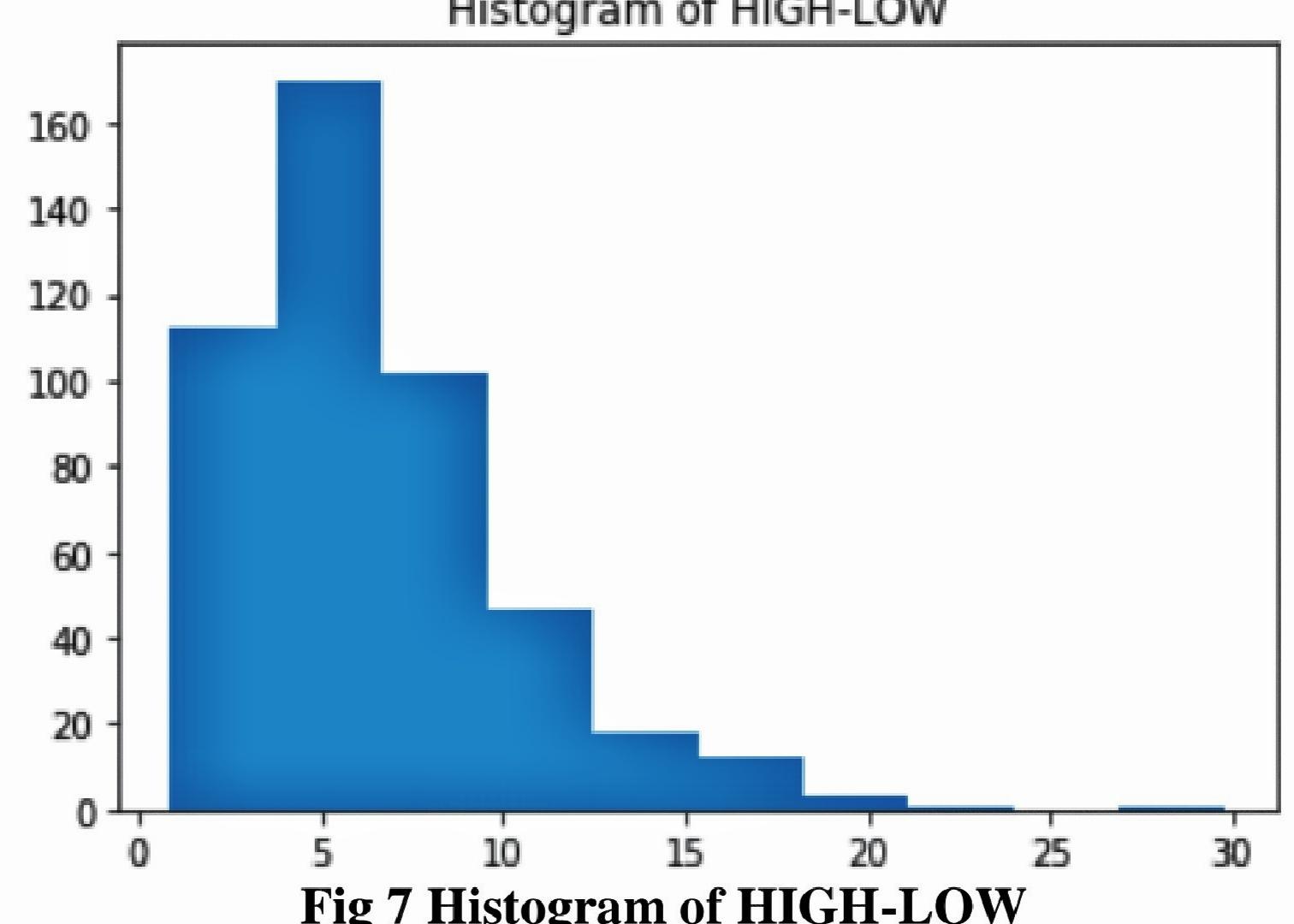
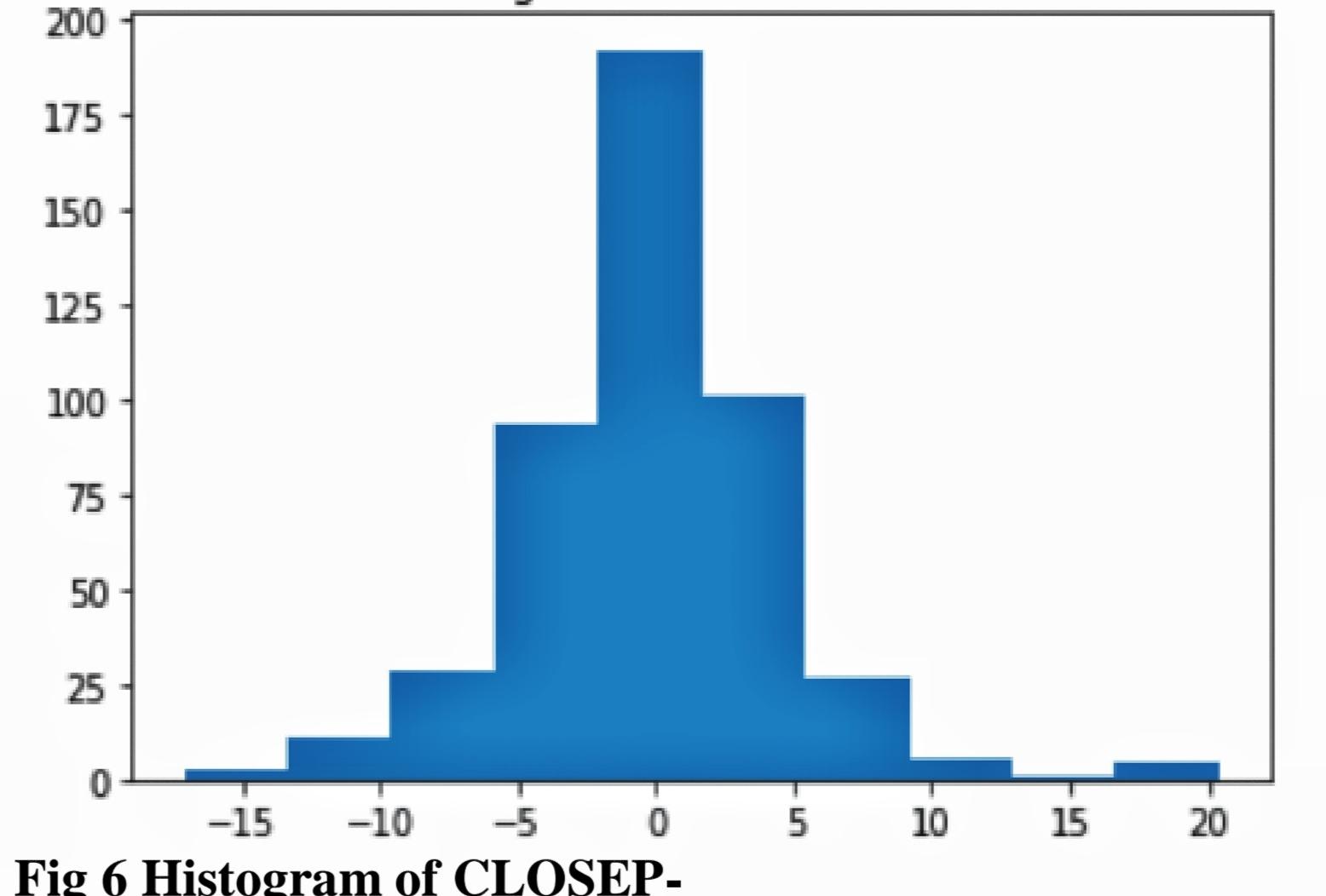


This is the result of using the head(). Since we are using the pandas library to analyse the data, it returns the first five rows. Here five is the default value of the number of rows it returns unless stated otherwise. The trading code in the processed data set is not relevant so we use the strip() to remove it and replace all of the trading codes with a value „GP‟

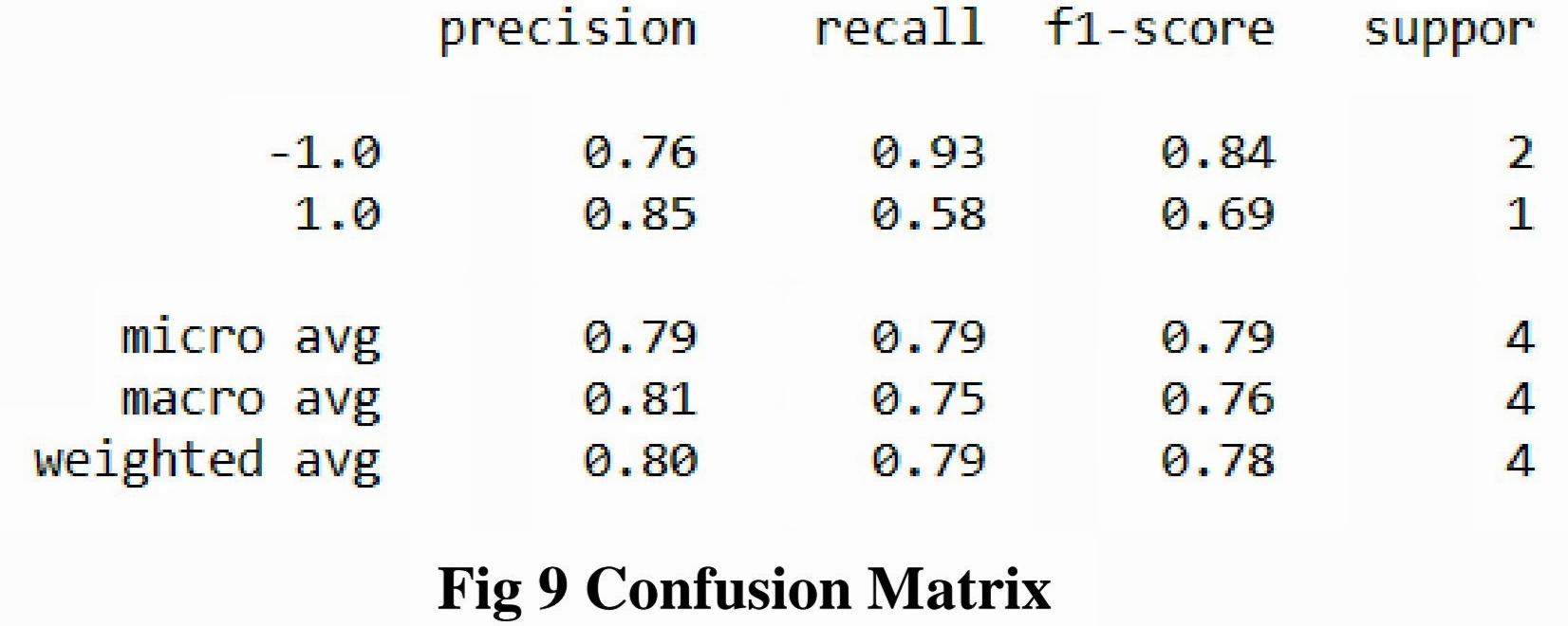


This is a time series plot generated from using the “matplotlib.pyplot” library. The plot is of the attributes “CLOSE UP” vs “DATE”. This is to show the trend of closing price of stock as time varies over a span of two years. The figure provided below is the candlestick plot, which was generated using the library “mpl\_finance”. The candlestick plot was generated using the attributes 'DATE', 'OPENP', 'HIGH', 'LOW','CLOSEP'.

##Fig 5,6,7##



The above two figures are histograms plotted between „CLOSED‟ and „OPEN‟ and the attributes „HIGH‟ and „LOW‟. This is done because we believe today‟s closing price and opening price along with the high and lowest price of the stock during last year will affect the price of the stock at a later date. Based on such reasoning we devised a logic “if today‟s CLOSEUP is greater than yesterday's CLOSEP then we assign the value 1 to DEX or else we assign the value -1 to DEX. Based on such the whole data set is processed and upon using the head() we get a glimpse of the data obtained thus far. The next step entailed the setting of feature and target variable, along with the setting of train size. Using the sklearn libraries we import the SVC classifier and fit it with the training data. After training the model with the data and running the test data through the trained model the confusion matrix obtained is shown below.



Along with this, we use the same dataset to train another model. This model utilises the Random Forest Classifier belonging to the ensemble technique. The decision trees have the default values so that leaves the “n\_estimator” value to be 10 since this is version 0.20. However, the value of “n\_estimator” will change to 100 in the version 0.22. After fitting the model with the data and running it against predicted data we find that this has an accuracy score of 0.808. To sum it up, the accuracy of the SVC Model in Test Set is 0.787 whereas the accuracy score of the random forest classifier is calculated to 0.808.

**CONCLUSION**

By measuring the accuracy of the different algorithms, we found that the most suitable algorithm for predicting the market price of a stock based on various data points from the historical data is the random forest algorithm. The algorithm will be a great asset for brokers and investors for investing money in the stock market since it is trained on a huge collection of historical data and has been chosen after being tested on a sample data.The project demonstrates the machine learning model to predict the stock value with more accuracy as compared to previously implemented machine learning models.

**FUTURE ENHANCEMENT**

Future scope of this project will involve adding more parameters and factors like the financial ratios, multiple instances, etc. The more the parameters are taken into account the more will be the accuracy. The algorithms can also be applied for analyzing the contents of public comments and thus determine patterns/relationships between the customer and the corporate employee. The use of traditional algorithms and data mining techniques can also help predict the corporation‟s performance structure as a whole.