[[1]](#footnote-1)

STOCK PRICE PREDICTION

E20CSE120-ARYAN SINGH CHAUDHARY,

E20CSE110-HARSH VERMA, E20CSE114-HARSH RAPEIRA

*Abstract*— Predicting stock market is very tough due to its volatile nature. It depends on geopolitics, the global economy, physical and psychological factors, and many more. Using artificial intelligence and machine learning we will try to predict the future value of stocks. Machine Learning and deep learning techniques will be used for making predictions in the stock market: - Random Forest- can predict the stock process. It uses techniques such as bootstrapping and pasting. It can randomly pick different features and create multiple Decision and Tree models. This is how it can help in predicting stocks. Artificial neural network: - Artificial Neural Network is a popular method that can also use technical analysis for prediction in stock markets. As the data for predicting the price of the stock can be an unstructured example, text-based news updates, blogs, etc. ANN can also deal with unstructured data.

# INTRODUCTION

T The stock market is known for being dynamic, unpredictable, and non-linear. Predicting stock prices is difficult because they are influenced by a variety of factors such as political conditions, the global economy, a company's financial reports, performance, and so on. Thus, strategies for predicting stock values in advance by examining the trend over the last several years could show to be quite valuable for making stock market moves to maximize profit and avoid losses. For predicting an organization's stock price, two basic methodologies have been offered in the past. For predicting the future price of a stock, the technical analysis method employs past stock prices such as closing and opening prices, the volume traded, adjacent close values, and so.

The second sort of study is qualitative, which is carried out by economic analysts using external factors such as firm profile, market situation, political and economic considerations, textual information in the form of financial news articles, social media, and even blogs. For predicting stock prices, advanced intelligent approaches based on either technical or fundamental analysis are now applied. The data size for stock market analysis is particularly large and non-linear. To deal with this variety of data, a model that can detect hidden patterns and complicated relationships in this vast data set is required. In this field, machine learning techniques have been shown to boost efficiencies by 60-86 percent when compared to previous methods.

*Motivation*

The problem of stock price prediction is a well-known and essential one. We can get insight into market behavior over time with an effective stock prediction model, recognizing tendencies that might otherwise go unnoticed. With the computer's increased computing capacity, machine learning will effectively overcome this challenge. However, many machine learning algorithms can't work with the public stock information , and adding extra characteristics can cost thousands of dollars per day. In this research, we will present a strategy for improving our findings by incorporating user predictions into the current machine learning algorithm while leveraging public historical data.

The underlying assumption is that if we had all of the information regarding today's stock trading (from all specific traders), the price will be predictable. As a result, even if we just have partial knowledge, we should be able to enhance the present prediction lot.

Getting daily user projections has become a practical task as the Internet, social networks, and online social activities have grown. As a result, our motivation is to provide a public service that incorporates historical data and user forecasts to create a more robust model that benefits everyone.

## Background Research

# Many significant changes have occurred in the financial markets environment over the last two decades. The development of strong communication and trading systems has broadened the range of options available to investors. Stock return forecasting is an important financial topic that has piqued the interest of researchers for many years. It is based on the notion that basic data that has been made publicly available in the past has some predictive value for future stock returns. Data mining techniques are innovative techniques that can be used to extract information from data in order to extract such associations from the given data. As a result, a number of researchers have concentrated on technical analysis and sophisticated math and science. Artificial intelligence and data mining approaches have gotten a lot of attention recently. Some models have been suggested and implemented using the approaches indicated above; for example, the authors of conducted an empirical investigation employing back propagation neural networks (BPNN) to build a stock buying/selling alert system; their NN was nicknamed NN5. From January 2004 to December 2005, the system was trained and evaluated using historical price data from Hong Kong and Shanghai Banking Corporation Holdings. The established system was able to anticipate short-term price movement directions with an accuracy of roughly 74%, according to the empirical results. The study used a decision tree technique to improve on Lin's work, which attempted to change the filtering rule of buying when the stock price rises k percent above its previous local low and selling when the stock price falls k percent from its previous local high. Combining three choice factors linked with basic analysis was presented as a modification to the filtering rule. Lin's strategy outperformed the filtering rule in an empirical test utilizing stocks of Taiwanese electronics businesses. According to Lin's research, the criteria for grouping trade points solely took into account previous data; future data was not taken into account at all.

# II RELATED WORK

For predicting stock prices, most previous work in this area has used classical algorithms such as linear regression Random Walk Theory (RWT) , Moving Average Convergence / Divergence (MACD) , as well as some linear models such as Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) . Machine learning has recently been shown to improve stock market prediction. Support Vector Machine (SVM) and Random Forest (RF) techniques . Some neural network-based techniques, such as Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and deep neural networks, such as Long Short Term Memory (LSTM), have also showed promise .

Through a self-learning method, ANN is capable of discovering hidden features. These are good approximators that can determine the input-output relationship of a huge, complex dataset. As a result, ANN is a strong choice for predicting an organization's stock price. Selvin et al used a comparative examination of different Deep learning approaches to predict the stock price of NSE listed businesses. Hamzaebi et al. used iterative and directive methods like the ANN model to experiment with multi-periodic stock market forecasting. Rout et al. used a low-complex RNN model to predict the stock market and tested it on the Bombay stock exchange and the S & P 500 index dataset .

Roman et al. trained RNN models on stock market data from five countries: Canada, Hong Kong, Japan, the United Kingdom, and the United States, and then utilised the networks to predict stock return trends . Yunus et al. used ANN on NASDAQ in 2014 to predict stock closing prices . Mizuno et al. used ANN to perform technical analysis on the TOPIX dataset and apply it to a system for predicting buying and selling timing . Random Forest (RF) has been proposed in some publications for predicting purposes. The technique of RF is an ensemble technique. In most cases, it can handle both regression and classification tasks. It works by training many decision trees, which then produce mean regression of individual decision trees.

On the New York electricity market, RF was used to accurately estimate real-time prices . Yand et al., for example, used an RF model to perform short-term load forecasting in the operation of electrical power systems . Herrera et al. employed RF as a predictive model for hourly urban water demand forecasts. In this study, two approaches, ANN and RF, were utilised to estimate an organization's closing price. The models employ a collection of new variables derived from a financial dataset that includes the Open, High, Low, and Close of a certain company. These new indicators will play a critical role in improving the accuracy of models in predicting a company's next day closing price. Two performance measures are used to assess the models' effectiveness: RMSE and MAPE.

Kim and Han developed a model for predicting stock price index that combined artificial neural networks (ANN) and genetic algorithms (GAs) with feature discretization. The technical indicators as well as the direction of change in the daily Korea stock price index were used in their research (KOSPI). They picked features and formulas from data containing samples of 2928 trade days spanning the years January 1989 to December 1998. They also used feature discretization optimization, which is a technique comparable to dimensionality reduction. The fact that they used GA to optimise the ANN is one of their work's merits.

First, the hidden layer's input characteristics and processing components are limited to 12 and are not changeable. Another disadvantage is that the authors only concentrated on two elements in the optimization phase during the ANN learning process. They still believe GA has a lot of potential when it comes to feature discretization optimization. Our initialised feature pool refers to the features that have been chosen. In , Qiu and Song presented a method based on an optimised artificial neural network model for predicting the direction of the Japanese stock market. The authors of this paper combine genetic algorithms with artificial neural network-based models to create a hybrid GA-ANN model.

Piramuthu evaluated and contrasted various feature selection strategies for data mining applications. He evaluated how different feature selection strategies enhanced decision tree performance using datasets such as credit approval data, loan defaults data, web traffic data, tam, and kiang data. The Bhattacharyya measure, the Matusita measure, the divergence measure, the Mahalanobis distance measure, and the Patrick-Fisher measure were among the probabilistic distance measures he compared. The Minkowski distance measure, city block distance measure, Euclidean distance measure, Chebychev distance measure, and nonlinear (Parzen and hyper-spherical kernel) distance measure are all inter-class distance measures. The author assessed both probabilistic distance-based and multiple inter-class feature selection methods, which is a strength of this paper.Furthermore, the author conducted the study using a variety of datasets, which added to the paper's strength. The evaluation algorithm, on the other hand, was solely a decision tree. We can't say whether the feature selection methods would hold up in a larger dataset or with a more complicated model.

In , Hassan and Nath used the Hidden Markov Model (HMM) to estimate stock market values for four major airlines. They divide the model's states into four categories: open price, close price, maximum price, and lowest price. The strength of this paper is that it does not require expert knowledge to construct a prediction model. While this research is limited to the airline industry and evaluated on a short dataset, it may not result in a generalizable prediction model. To accomplish the comparison job, one of the methodologies used in stock market prediction-related activities could be used.

# To forecast stock price trends, Lei used a Wavelet Neural Network (WNN). As an optimization, the author used Rough Set (RS) for attribute reduction. The stock price trend feature dimensions were reduced using Rough Set. It was also used to identify the Wavelet Neural Network's structure. This study's dataset includes five well-known stock market indices: (1) the SSE Composite Index (China), (2) the CSI 300 Index (China), (3) the All Ordinaries Index (Australian), (4) the Nikkei 225 Index (Japan), and (5) the Dow Jones Index (USA). The model was evaluated using various stock market indexes, and the results were convincing in terms of generality. The computational complexity is reduced by employing Rough Set to optimize the feature dimension before processing. However, in the discussion section, the author mainly emphasized parameter tweaking and did not mention the model's flaws. Meanwhile, we discovered that while the evaluations were done on indices, the same model would not perform as well if applied to an individual stock.

Weng et al. used ensemble methods of four well-known machine learning models to forecast short-term stock prices. There are five sets of data in this study's dataset. These datasets were gathered using three open-source APIs and the TTR R package. (1) neural network regression ensemble (NNRE), (2) a Random Forest with unpruned regression trees as base learners (RFR), (3) AdaBoost with unpruned regression trees as base learners (BRT), and (4) a support vector regression ensemble were the machine learning models they utilised (SVRE).

A detailed examination of ensemble approaches for predicting short-term stock prices. The authors chose eight technical indicators in this study based on their background knowledge and then carefully evaluated five datasets. The main contribution of this article is that it established an R-based platform for investors that does not require users to submit their own data but instead calls an API to retrieve data from an internet source. From a research standpoint, they only looked at price predictions for 1 to 10 days ahead of time, but not for longer than two trading weeks or for shorter than one day.

The model may not be generalised to other stock markets or require further revalidation to check if it suffers from overfitting concerns due to the fact that they only looked at 20 U.S.-based stocks.

Thakur and Kumar also used multi-category classifiers and random forests to create a hybrid financial trading support system (RAF). They used stock indices from the NASDAQ, DOW JONES, S&P 500, NIFTY 50, and NIFTY BANK to perform their research. To create "Buy/Hold/Sell" signals, the authors suggested a hybrid model that integrated random forest (RF) techniques with a weighted multicategory generalised eigenvalue support vector machine (WMGEPSVM). They employed Random Forest (RF) for feature trimming before analysing the data. The authors developed a viable model for real-world investment operations that may create three fundamental signals for investors to consider. They also compared and contrasted a number of comparable methods. They did not, however, indicate the length of time or computing complexity of their works.

Meanwhile, the absence of financial domain knowledge was an unavoidable difficulty in their work. Investors consider indices data to be one of the features, but they are unable to use indices to directly run a specific stock.

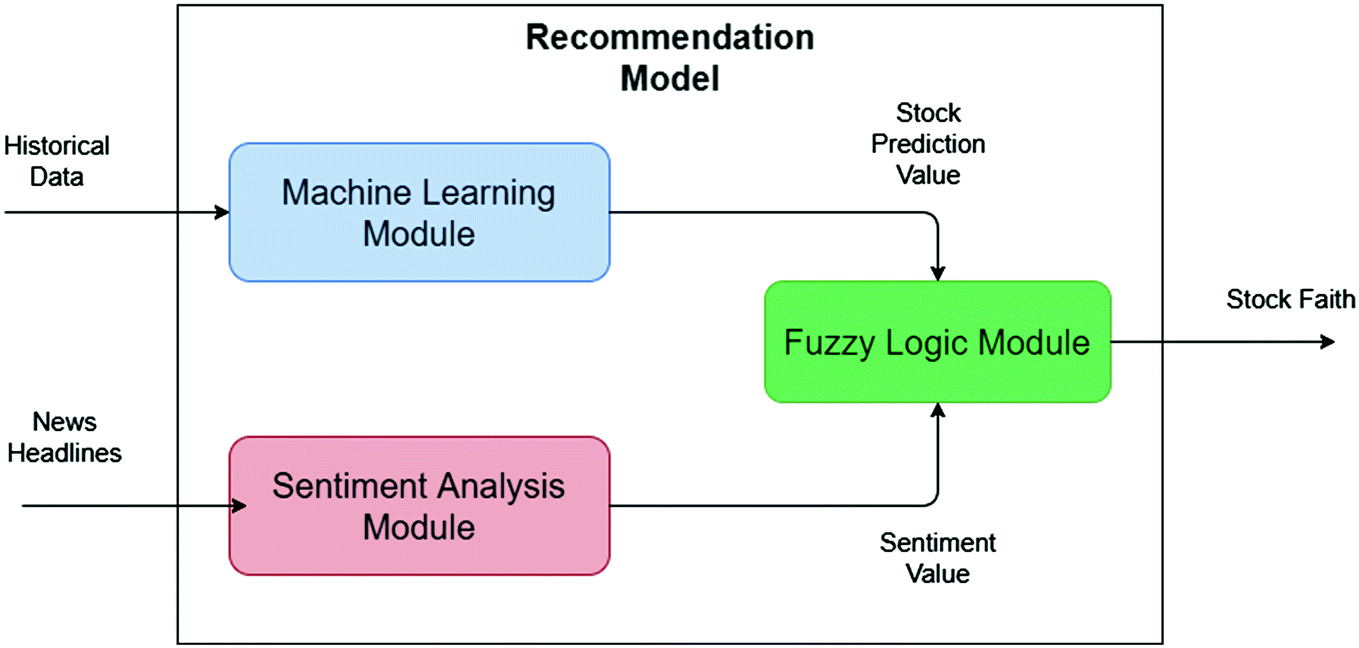
Nekoeiqachkanloo et al. suggested a stock investment system with two different ways. The advantages of their proposed solution are self-evident. To begin with, it is a complete system that includes data pre-processing as well as two separate algorithms for recommending the finest investment segments. Second, the system includes a forecasting component that preserves the time series' characteristics. Finally, their input features are a combination of fundamental and technical indicators that try to bridge the gap between the financial and technical domains.

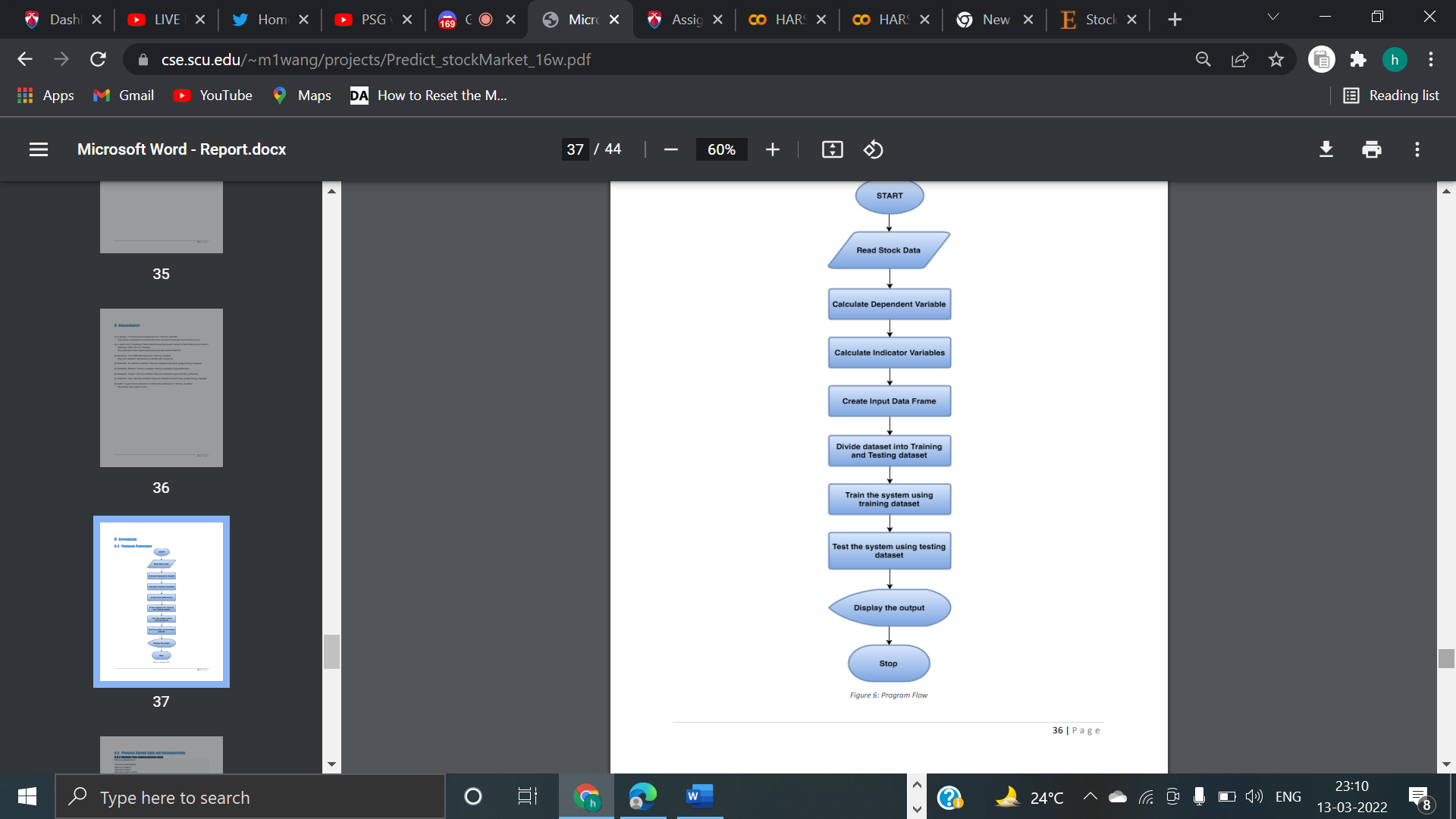
Their study, however, has a flaw in the evaluation section. They chose 25 well-known equities to test the suggested technique instead of a big dataset. There's a good chance that the well-known stocks have some hidden characteristics in common.

Limited data-preprocessing techniques established and used is one of the key flaws observed in similar efforts. The majority of technical work is on developing prediction models. When they choose the features, they make a list of all the features that have been discussed in earlier works, run them through the feature selection algorithm, and then choose the features that have received the most votes. Similar works in the investing realm have demonstrated a greater interest in behaviour analysis, such as how herding habits affect stock performance or how the percentage of inside directors who own the firm's common stock affects stock performance. To spot these patterns, established technical indices and investment experience are frequently used as a pre-processing procedure.

In related research, rather than doing feature choices, a full statistical analysis is often performed based on a special dataset to conclude new features. Some information, such as the percentage of an index's variation, has been shown to influence stock performance. We believe that extracting new features from data and integrating them with current common technical indices will boost existing and well-tested prediction models greatly.

# III BLOCK DIAGRAM





IV. PROPOSED METHODOLOGY

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 algorithm, 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 random forest classifier which has the same hyper parameters as of a decision tree.

V EXPERIMENTAL SETUP

**Pre-processing**

Diagram

Description automatically generated

Diag-: Data Pre-processing

Every dataset consists of various types of anomalies such as missing values, redundancy, or any other problem for removing this problem there is need of certain step called as processing data. Pre-processing step is needed to overcome from such problem. There are three preprocessing steps:

**Formatting:** The data set is used for implementation is taken from Kaggle; it may contain certain attributes whose names are not clear in the (dataset name) also contain certain unrelated attribute which is not useful for the greater performance of proposed work.

**Cleaning:** Pre-processing or cleaning means is to remove or fixing of missing out entry in the data frame. Row containing these incomplete columned to be removed also for removing certain redundant entries in data frame this step is recommend

Sampling: Sampling is also done on the dataset to enhance the performance of the algorithm on sample data set may lead algorithm to take longer time.

**Feature Extraction**

Closing price is being predicted in our work. So, the features that contributes much to the prediction of closing price is extracted in this step. Then, the features are given as input to MACD (Moving Average Convergence/Divergence Oscillator). Followed by this the features are passed on to recurrent neural network algorithm.

Diagram

Description automatically generated

1. [↑](#footnote-ref-1)