International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

# STOCK MARKET PREDICTION: AN ENSEMBLE APPROACH

**Cecil Ignatius Hermina[1], Sreelakshmi Varma M[2], Dr. N Kanthimathi[3], Saranya N[4]**

[1] *Department of Electronics and Communication, Bannari Amman Institute of Technology*
[2] *Department of Electronics and Communication, Bannari Amman Institute of Technology*
[3] *Department of Electronics and Communication, Bannari Amman Institute of Technology*
[4] *Department of Electronics and Communication, Bannari Amman Institute of Technology*

---

**Abstract -** AI stock market prediction requires processing a lot of data and making predictions based on that analysis, which is a challenging task. By analysing data from a variety of sources, including news stories, earnings reports, and social media, AI techniques like machine learning algorithms and natural language processing can help discover trends and forecast changes in the stock market. It is crucial to keep in mind, though, that numerous unknown events have an impact on the stock market. For investors, forecasting is highly challenging to do profitably because of the noisy and static data. It is advised to employ AI predictions as one of several inputs in a well-diversified investment plan to attain accurate stock market predictions. The most recent stock market-related forecast methodology, despite considerable effort, contains major flaws. It makes sense to assume that this study's stock prediction is the result of an integrated process.

***Key Words***: Stock prediction, ML, AI, LSTM, Linear regression, SVC.

## 1.INTRODUCTION

### 1.1 STOCK MARKET

On the stock market, one can buy and sell shares of companies that are openly traded. The stocks, which are also referred to as assets, reflect ownership in the business. To make the sale and purchase of shares possible, the stock market serves as a middleman. Machine learning-based share market analysis can be used to determine the current value of business shares and other investments that are traded on an exchange.

The development of stock market prediction has become increasingly important to professional analysts and investors. Due to the chaotic environment in the market, it is extremely challenging to analyze price and stock market movements. The release of quarterly earnings releases and market news are just two of the many factors that influence stock price changes. Stock market indices are developed using the market capitalization of the companies. So, in the world of markets, it is quite challenging to make accurate stock market predictions. The development and testing of stock market behavior has drawn the attention of scholars and market professionals.

### 1.2 STOCK MARKET FORECASTING:

The stock market trader declares their intention to make money by making an investment on the stock market. Innovative applications that could make forecasting the stock market lucrative have piqued investor interest in the stock market. Stock market forecasts that are accurate depending on prior knowledge. The ability to watch and manage the market with the help of stock market forecasting tools enables users to take the proper actions. The various industrial stock information components, which affect the entire financial market, must be handled by the stock market. They are modified in accordance with the business conditions of investors who consider sales and purchases. Future estimated revenue, a news report on earnings, management changes, etc. are just a few things that might alter the market's position. Therefore, precise stock market forecasts assist investors in making decisions. Investors can learn more about the stock market and make money by using ML techniques.

### 1.3 MACHINE LEARNING:

A wide range of models and algorithms that can be used in a variety of use cases fuel predictive analytics tools. To get the most out of a predictive analytics system and use data to make wise decisions, the most effective predictive modelling approaches must be found. Machine learning contains concepts for automation, but it still requires human oversight. Artificial intelligence is used in machine learning, which uses available data to handle or assist in handling statistical data using algorithms. To create a machine learning system that excels on data examples that have never been seen before, a high degree of generalization is required.

Computer science's relatively new field of study, machine learning, provides a variety of data analysis techniques. While many other methods are not based on well-known statistical techniques, principal component analysis and logistic regression are good instances. Most statistical methods function by choosing the probabilistic model from a set of linked models that best fits the observed data. Machine learning methods generally aim to find models that best-fit data, with the exception that they are no longer restricted to probabilistic ones (i.e., they solve specific optimization problems). This is comparable to conventional refining methods.

Therefore, machine learning methods are superior to statistical ones because they do not necessitate underpinning probabilistic models, in contrast to the latter. Even though some machine learning techniques use probabilistic models, because of the complexity and variety of data sources, conventional statistical techniques are frequently inappropriate for the coming big data era. Predicting probabilistic models with variables from various data sources that are simultaneously reliable and amenable to statistical analysis may be exceedingly difficult, if not impossible.
A wider class of alternative analysis techniques that are more adaptable and better suited to contemporary data sources may be offered by machine learning. Statistical agencies must

investigate the potential applications of different methods for machine learning to see if they can meet their future needs more effectively than with more conventional approaches.

*1.3.1 Classes of Machine Learning*

*Examples of supervised learning:*

Logistic regression is an example of a regulated machine learning technique used for prediction. In logistic regression, a limited number of observation units are used to monitor a binary response variable (say, having a value of 0 or 1) and a number of predictor variables (covariates). This is referred to as the acquired data in terms of machine learning. The basic hypotheses are that the correlation between both the predictor variables and the response variables is such that the response variable has a Bernoulli distribution, a type of probability model, as well as the posterior probability of the response's logarithm are a linear function of the predictors. The units' response variables are taken to be independent of each other, and their combined probability distribution is made subject to the maximum likelihood technique to determine the best values for the coefficients that parameterize the above said joint distribution in this linear function. A particular model with these ideal coefficient values is referred to as a "fitted model," and it can be used to "guess" the response variable values for new units for which only the predictor values are known or to "classify" the new units as 0 or 1. One non-statistically supervised machine learning method is Support Vector Machines (SVM). They aim to select the SVM model that best matches training data and then use it to make classifications, much like the logistic regression classifier just mentioned.

*1.3.2 Among the most popular predictive algorithms is*

*Decision trees*: Decision trees are an effective and easy-to-use technique for numerous variable analyses. They are created by algorithms that discover different methods of segmenting data into branches. Based on groups of input factors, decision trees divide data into subsets.

*Regression (linear and logistic):* One of the most widely used statistical techniques is Regression. With large and varied data sets, regression analysis evaluates the relationships between variables, discovering significant patterns and how they are related to one another.

*Neural networks:* Artificial neural networks, also known as neural networks, are a type of deep learning algorithms that mimic the actions of neurons in the human brain. They are incredibly helpful for analyzing huge data sets and are frequently used to solve difficult pattern identification issues. They work well when a few of the variables are unknown and are excellent at managing nonlinear relationships in data.

*1.3.3 Developing the Right Environment:*

Although machine learning and predictive analytics can be a huge help to any organization, installing them carelessly without considering how they will integrate into daily operations would severely limit their potential to provide the insights the organization needs.

*1.3.4 Understanding Predictive Models:*

Data scientists and IT professionals in an organization are usually in charge of selecting the best prediction model or creating their own to suit the organization's requirements. Mathematicians, researchers, data scientists, and business analysts are now the only professionals with expertise in predictive analytics as well as machine learning. Companies' employees are using it more frequently to learn new things and improve corporate operations, but issues might occur when they are unsure of the best model to apply, how to put it into practice, or when they need information urgently.
SAS develops sophisticated tools to help companies with analytics and data governance. Within the same environment, the data governance tools help companies maintain high-quality data, coordinate business operations, and pinpoint data issues. The predictive analytics solution is designed to help companies turn their analytics into immediate knowledge and insight for better, more effective decision-making. All users' needs are met by these predictive analytics tools, which also enable fast deployment of these predictive models.

1.4 WEB SCRAPPING:

"Web scraping" is a technique for automatically gathering massive quantities of website data. The bulk of this data is in HTML, that is an unstructured format that must be converted into spreadsheets or database containing organized data in order to be used in various applications. Web scraping is a technique that can be used in many various ways to gather data from websites. These include the use of specific tools, online services, or even the creation of your own custom web scraping code. The APIs of many well-known websites, such as Google, Twitter, Facebook, Stack Overflow, and others, can also be used to retrieve structured data. This is the best option, but due to a lack of technical know-how or because they choose not to, some websites do not permit users to obtain significant amounts of structured data. In these circumstances, it is recommended to use web scraping to collect data.

For a web scraper to scrape a website, the URLs must first be given. This completes the loading of the HTML content for those websites. A more sophisticated scraper may be able to fully retrieve the JavaScript and CSS components. The crawler then extracts the pertinent information from this HTML code and outputs it in the manner selected by the user. The data can be saved in a variety of formats, including a JSON file, though it is usually saved as an Excel spreadsheet or a CSV file.

*1.4.1 What is Web Scraping used for?*

Web Scraping has different uses in multiple industries. A few of them are listed below:
*i) Price Monitoring:* Businesses can use web scraping to collect product information for both their own and similar competing products to evaluate how it affects their pricing strategy. Businesses can use this information to determine the best price for their items in order to maximize revenue.
*ii) Market Research:* Market research can greatly benefit from web scraping. Large volumes of high-quality web-scraped data can be quite beneficial for businesses in analyzing consumer behaviors and determining the approach they should adopt in the future.

*iii) News Monitoring:* A company or organization can get in-depth information on the most recent news by scraping several news websites. For organizations that often appear in the media or rely on the daily news for everyday operations, this is even more crucial.
*iv) Sentiment Analysis:* Sentiment analysis is necessary if businesses wish to comprehend how consumers generally feel about their products. Web scraping is technique businesses use to gather information from social media websites like Twitter and Facebook to assess the public's opinion of their products. As a result, they will be able to surpass their competitors and create products that people will demand.
*v) Email Marketing:* Web scraping is another tool that businesses may use for email marketing. Web scraping allows them to gather Email IDs from numerous websites, and they may then send mass promotional and marketing emails to everyone who has one of these Email IDs.

## 2. LITERATURE SURVEY

Ernest Kwame Ampomah et al (2020) studied that the predictive model is built using the combination of the GNB algorithm. The research seeks to close the gap in the literature regarding the capability of the Gaussian Naive Bayes ML algorithm to predict stock price movement by examining the functioning of the GNB algorithm when integrated with various feature scaling and feature extraction methods. Using Kendall's test of concordance for the various evaluation parameters, the setup GNB models were rated for effectiveness. The results demonstrated that the GNB LDA predictive model, which unifies the GNB algorithm and linear discriminant analysis, outperformed all other GNB models [1].

M. J. Awan et al (2020) studied multiple machine learning models are developed using PySpark and Spark MLlib, which have greater performance than conventional models and are scalable, quick, and easy to relate to other tools. With the use of MLlib models such as decision trees, generalized linear regression, random forests, and linear regression, the historical stock prices of the top ten companies were examined. Classification models like logistic regression and Naive Bayes were used. According to experimental findings, linear regression, random forest, and generalized linear regression offered better accuracy. The decision tree's experimental findings did not accurately predict changes in share prices on the stock market [2].

Omar D. Madeeh et al (2020) researched effective machine learning techniques to build a solid model for stock market prediction. The study was divided into three stages first, gathering the data set of the stock market; second, applying K-Nearest Neighbor (K-NN) and Random Forest (RF), two supervised machine learning approaches; and third, evaluation, which assessed the effectiveness and precision of the predictions for the two suggested models [3].

Pooja Mehta et al (2020) suggested an algorithm that considers the sentiment of the public, opinions, news, and previous stock prices to forecast future stock values. Using well-known social networking sites and individual blogs, people may actively share their opinions and suggestions for a variety of goods and services. Opinion-sharing websites include social networking sites like Twitter, Facebook, and Google+. The stock market (SM) is an essential element of businesses and the economy and has a significant effect on the expansion of business and industry. It is well known that academics are interested in predicting trends in social media. Social networking is an ideal way to capture how the public feels about current events. Stock trend prices are thought to be affected by financial news, as well as several data mining methods are used to handle fluctuations in social media. When dealing with predictions relating to social media, machine learning can offer a more reliable and accurate method. This study sought to establish a link between changes in a company's stock price and publicly accessible sentiments (opinions) about it. When creating and putting into use a stock price prediction accuracy instrument, the general public's mood was taken into consideration in addition to other elements [4].

Salvatore M. Carta et al (2020) made a machine learning approach that is used to solve the binary categorization problem, which aims to predict the weight (high or low) of impending share price changes of each company in the S&P 500 index. With the aim of finding the most popular keywords on the market during a particular period of time and within a specific industry sector, sets of lexicons are created from internationally distributed publications. Following feature engineering on the generated lexicons, a Decision Tree classifier is given the resulting features to process. The expected label (high or low) indicates how the stock price of the underlying company will change the following day, either rising or falling above or below a predetermined threshold. This technique outperforms the opposition, according to the performance assessment that was done using a walk-forward technique and a set of reliable baselines [5].

## 3. METHODOLOGY

3.1 PROPOSED SOLUTION:

Step 1: Data Collection:
Using web scraping, historical data is gathered.

Step 2: Data Pre-processing:
2.1. Data discretization: Data reduction
2.2. Data transformation: Normalization
2.3. Data cleaning: Fill in the missing values.
2.4. Data integration: Integration of data files

Step 3: Feature Extraction:

Necessary features from date, open, high, low, close, and volume are chosen.

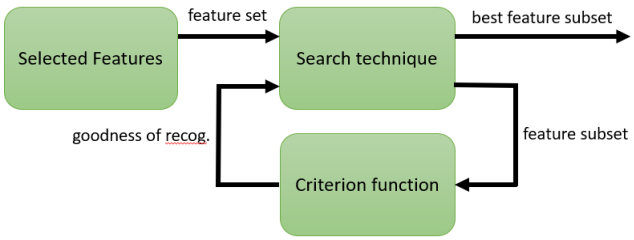


Fig 3.1. Block diagram of feature extraction

Step 4: Training Neural Network:

Data is fed to the neural network and trained for predictions. Web scraping for the stock price is automatically extracting stock price data from websites that provide such information. This technique involves using web scraping tools to collect stock price data from different financial websites, and then processing the data to obtain insights for investment decisions. Web scraping for stock price involves accessing websites that provide stock data, such as Yahoo Finance, and Google Finance, and extracting the stock price data from these sites. This data can include information on the stock's opening and closing prices, trading volume, stock charts, financial ratios, news, and other relevant information.

Once the data is collected, it can be analysed using statistical and machine learning techniques to identify trends, patterns, and relationships that can inform investment decisions. For example, analysts can use web-scraped data to track changes in stock prices, identify significant market events, and predict future price movements.

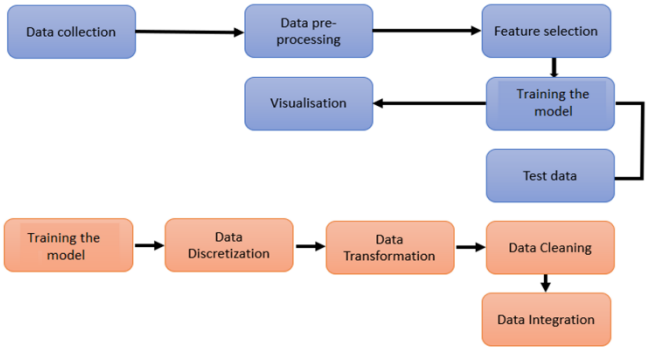


Fig 3.2. Block diagram of proposed methodology

3.2. ALGORITHM:

3.2.1. Stacking:

Stacking is an ensemble learning method used in machine learning to combine the predictions of various models to create a prediction that is more accurate. Multiple models are trained on the same dataset during stacking, and the forecasts from each model are then added together using a meta-model.

Using multiple models' predictions as input to a higher-level model that learns how to combine them to create a final prediction is known as stacking. A meta-model, also known as a blender, is a higher-level model that is trained using the predictions from the basic models to discover the best way to combine them.

## 4. RESULTS AND DISCUSSION:

4.1 LINEAR REGRESSION:

This is used to find the state between the independent and dependent labels.
a = nb + e (where a and b are dependent or independent labels).
Fig 4.1 illustrates the actual vs predicted values using linear regression.

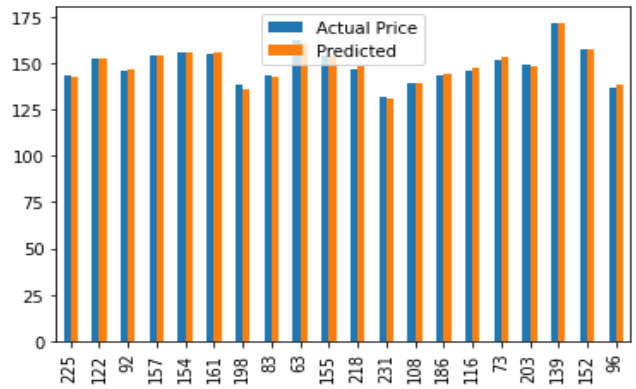


Fig 4.1. Actual vs Predicted values using Linear Regression

4.2 LSTM (Long short-term memory):

Since LSTMs can learn long-term dependencies between data time steps, they are primarily used to learn, process, and categorize sequential data. Fig 4.2 illustrates the actual vs predicted values using LSTM.

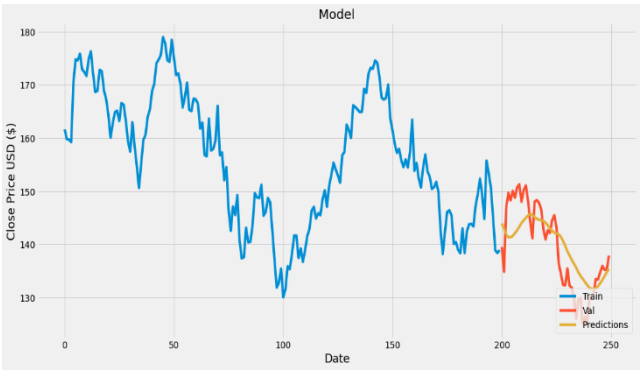


Fig 4.2. Actual vs Predicted values using LSTM.

4.3 SUPPORT VECTOR CLASSIFIER:

The objective of a Linear SVC is to group the data you provide into categories using the "best fit" higher dimensional space that it gives. Fig 4.3 illustrates the actual vs predicted values using SVC.


Fig 4.3. Actual vs Predicted values using SVC.

4.4 STACKING:

By combining the strengths of various models, stacking has the potential to increase prediction precision, which is one of

its primary benefits. Stacking is versatile and can be applied to a broad range of machine learning models and algorithms. To successfully train the meta-model, stacking can, however, be computationally expensive and necessitate more data. Fig 4.4 illustrates the actual vs predicted values using stacking.

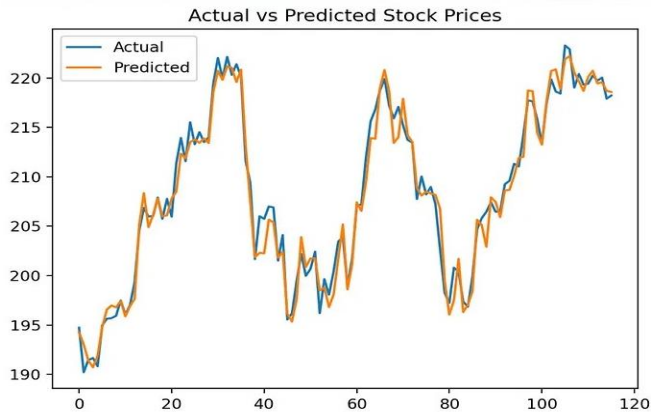

Fig 4.4. Actual vs Predicted values using Stacking.

**5. CONCLUSION:**

The work demonstrated the potential use of machine learning in analyzing the stock market based on the company name, previous price, and current prices. The accuracy of the predictions is above 85 percent. This survey's goal is to categorize current methodologies for using different datasets, performance matrices, and applying techniques. It makes use of 30 research articles from the most prestigious journals. The stock market prediction techniques are categorized using various ML algorithms. To improve prediction accuracy, some of the selected studies use hybrid methods in the stock market. The most critical step in forecasting stock markets is that the stock market is volatile. The market goes up and down and it's not always easy to predict. The dropping prices are affected because of the imbalance between supply and demand, interest rates, political factors, natural calamities, and inflation.

**6. REFERENCES:**

[1] Ernest Kwame Ampomah, Gabriel Nyame, Zhiguang Qin, Prince Clement Addo, Enoch Opanin Gyamfi and Michael Gyan, “Stock Market Prediction with Gaussian Naïve Bayes Machine Learning Algorithm”, 2020.
[2] Omar D. Madeeh and Hasanen S. Abdullah, “An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market”, 2020.
[3] M. J. Awan, M. Shafry, H. Nobanee, A. Munawar and A. Yasin, “Social Media and Stock Market Prediction: A Big Data Approach”, 2020.
[4] Salvatore M. Carta, Sergio Consoli, Luca Piras and Alessandro Sebastian Podda, “Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting”, 2020.
[5] Pooja Mehta, Sharnil Pandya and Ketan Kotecha, “Harvesting social media sentiment analysis to enhance stock market prediction using deep learning”, 2020.