International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Stock Market Prediction and analysis Using Machine Learning

P.YUGANDHAR

GMRIT CSE Rajam, India
yugandhar248.pilla@gmail.com

ABSTRACT :

Stock market prediction using machine learning explores the challenges of forecasting stock prices in a volatile market, where traditional methods like technical and fundamental analysis have shown inconsistent results. It emphasizes the implementation of a Random Forest approach, which is effective in predicting stock prices and returns. The study highlights the significance of selecting relevant parameters that impact share prices and incorporates sentiment analysis to evaluate the polarity of news articles, volumes, previous performance, 52 weeks high and low for enhancing the prediction accuracy. While analyzing different Machine Learning Algorithms for better accuracy, the proposed paper demonstrates the better efficiency of the Random Forest method in providing reliable forecasts for the benefit of investors

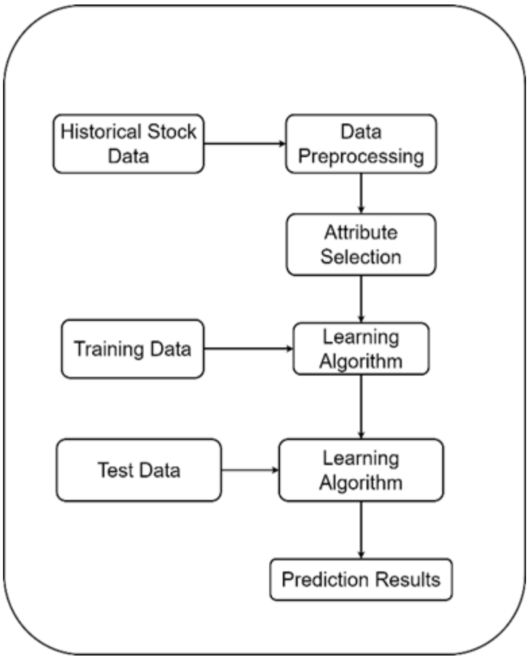
Keywords—Machine Learning, Stock market prediction, Random Forest, Sentiment analysis**.**

I. INTRODUCTION :

In the paper "Stock Market Prediction Using Machine Learning, " that was written by Subasi et al. in 2021, various machine learning algorithms were explored to predict stock market trends. Predicting stock market trends is intrinsically complex and demanding because the price and movement of the stocks have nothing in common with predictability. It is widely influenced by economic conditions, political events, investor sentiments, etc. A comparative study has been developed on the performance of seven different classifiers based on the machine learning approach: Random Forest, Bagging, Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Artificial Neural Networks (ANN). The classifiers were tested for accuracy improvements against normal datasets as well as datasets including leaked information. This paper demonstrates the two methods, Random Forest and Bagging, as the best-ofbreed-performer, but especially while training on leaked data. Therefore, this comparative study is a contribution to the continued research into improved forecasting of the stock market using machine learning methods.

The study highlights that both Random Forest and Bagging outperformed the other classifiers, especially on datasets with leaked information, showcasing their robustness in enhancing stock market forecasting accuracy. This comparative analysis underscores the significance of continued research into machine learning techniques for more effective and accurate stock market predictions, paving the way for further advancements in this field. The findings indicate the potential of ensemble methods, particularly Random Forest and Bagging, as reliable predictors in financial forecasting applications.

Figure 1. Architecture of stock market price prediction



The Stock Market is the accumulation of stockbrokers, traders, and investors who sell buy or share trades. There are so many companies that provide their stock list on market, these make their stocks attractive to investors [1]. Because ever since the 16s investors are trying different techniques to get knowledge about different companies to improve their investment returns [2]. It plays a very important role in increasing a developing country's economic status like India
[3].
The demand for Stock Market is growing significantly. We all know that it has been in focus for many years because of the outstanding profits [4]. Lots of wealth are traded daily through the stock market and so it is seen as one of the most profitable financial outlets [5]. Now, the stock market is one of the factors which shows a country's economy [6]. Many people invest a handsome amount of money in the share market but sometimes they tend to incur very huge losses because they depend upon the stockbrokers, who advise investors based on fundamental, technical, and time series [7]. Investors have been trying to find an intelligent idea to overcome such problems. This is where Stock Price Prediction comes into action because predicting stock prices is very necessary [2].

LITERATURE SURVEY :

Subasi et al. (2021) applied machine learning algorithms to forecast stock market trends, aiming to improve the accuracy and reliability of stock price predictions [1]. Strader et al. (2020) offered a comprehensive review of machine learning methods in stock prediction, identifying gaps and suggesting future research directions in predictive accuracy and model efficiency [2]. Similarly, Bansal et al. (2022) explored advanced machine learning techniques to achieve higher prediction accuracy, focusing on optimized algorithms for real-time forecasting [3].
Vijh et al. (2020) conducted an empirical analysis using various machine learning techniques to predict stock closing prices, demonstrating significant improvements in forecasting precision through hybrid models [4]. Kompella and Chakravarthy (2020) evaluated machine learning methods for stock market prediction, examining the strengths and limitations of each approach and suggesting combinations for robust predictions [5]. Li and Pan (2022) proposed an ensemble deep learning model that integrates stock prices with news data, showing promising results in capturing market sentiments for more accurate forecasts [6].Nti et al. (2020) reviewed fundamental and technical analyses used in stock prediction, highlighting how machine learning can enhance these traditional approaches [7]. Nabipour et al. (2020) compared continuous and binary data approaches for predicting stock trends using machine learning and deep learning, emphasizing the utility of a hybrid model for dynamic market conditions [8]. Shen and Shafiq (2020) developed a comprehensive deep learning model for shortterm stock trend prediction, achieving high accuracy through an adaptive learning mechanism [9].Kumbure et al. (2022) provided a literature review on machine learning techniques and data requirements for stock forecasting, underscoring the need for quality data to boost predictive accuracy [10]. Li and Bastos (2020) systematically reviewed deep learning models in technical analysis, highlighting the model structures best suited for volatile stock markets [11]. Zhang et al. (2019) introduced a generative adversarial network (GAN) for stock market prediction, showcasing how GANs can model complex market patterns effectively [12].Mehtab and Sen (2022) compared machine learning and deep learning algorithms for stock prediction, finding that deep learning often outperforms traditional models, especially with large datasets [13]. Rezaei et al. (2021) applied deep learning with frequency decomposition for stock price prediction, demonstrating enhanced forecasting accuracy through spectral analysis [14]. Bhattacharjee and Bhattacharja (2019) compared traditional statistical methods with machine learning for stock price prediction, establishing the superior accuracy of machine learning techniques [15].Asad (2015) optimized stock predictions through ensemble learning, demonstrating how combining multiple models can improve predictive outcomes [16]. Yoo et al. (2005) surveyed machine learning techniques using event-based information for stock prediction, emphasizing the importance of contextual market events in model design [17]. Gupta and Dhingra (2012) explored hidden Markov models for stock market prediction, identifying significant patterns in price movement trends [18].
Reddy and Sai (2018) demonstrated the effectiveness of machine learning in stock market forecasting by comparing different model approaches and identifying optimal settings for high accuracy [19]. Usmani et al. (2016) examined various machine learning techniques for stock market prediction, focusing on their performance in dynamic market environments [20]. Jiang (2021) reviewed recent advances in deep learning for stock prediction, underscoring the potential of novel architectures in handling complex market data [21]. The remainder of the paper is organized in the following sections. First, the process used to identify relevant studies is described. Next, based on an assessment of the studies identified, a research framework (taxonomy) is presented that groups the studies based on the ML technique used to predict stock market index values and trends. The studies in each category are then individually summarized and discussed to identify common findings, unique findings, limitations, and areas where more study is needed. The final section provides some answers related to the overall study objective that focuses on identifying directions for future research and recommendations for improving study generalizability.

METHODOLOGY:

Stock price prediction is an inherently complex and dynamic task due to the stochastic, volatile, and non-linear nature of financial markets. The unpredictable interplay of various macroeconomic, geopolitical, and companyspecific factors introduces significant challenges in accurately forecasting stock movements. Traditional statistical methods often fall short in capturing the intricate and multi-dimensional relationships present in financial data. This has led to the adoption of machine learning algorithms, which have demonstrated remarkable potential in uncovering hidden patterns and extracting valuable insights from large, complex datasets.
Machine learning algorithms excel in their ability to learn from historical data and adapt to evolving trends, making them well-suited for stock price prediction. By leveraging a combination of data-driven learning and advanced computational techniques, these models are capable of identifying subtle correlations and dependencies that may not be apparent through traditional approaches. Furthermore, machine learning models can integrate diverse financial features such as historical stock prices, trading volumes, technical indicators, and even external variables like news sentiment and economic indicators, thus providing a holistic approach to stock market analysis.

This study employs a range of machine learning methodologies, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Random

Forests, and Decision Trees, to explore their effectiveness in predicting stock prices. These methodologies were chosen due to their complementary strengths in handling different aspects of financial data:

1)Artificial Neural Networks (ANN):

An Artificial Neural Network (ANN) is a computational system inspired by the structure and functioning of biological neurons in the human brain, although not identical. ANNs have become widely used in various domains, including stock price prediction, over the past few decades. These networks consist of interconnected nodes, organized into three primary layers: the input layer, hidden layers, and the output layer. The input layer serves as the interface between the user and the network, accepting various input variables such as historical stock prices, trading volumes, and technical indicators. These inputs are then passed into the hidden layers, located between the input and output layers. The hidden layers are responsible for identifying intricate patterns and extracting meaningful features from the input data, leveraging weighted connections and activation functions to enhance learning. Finally, the processed information reaches the output layer, which generates the final prediction, such as a stock price forecast or classification result.

The ANN processes data by multiplying each input by a specific weight, representing the importance of that input, and applying an activation function to determine whether the neuron should activate. The activation function introduces non-linearity, enabling the network to model complex relationships within the data. Through repeated adjustments of weights during training, the network minimizes prediction errors and improves its accuracy. This layered structure and learning mechanism allow ANNs to capture hidden dependencies and trends in stock market data, making them a powerful tool for forecasting and decision-making in financial analysis**.**

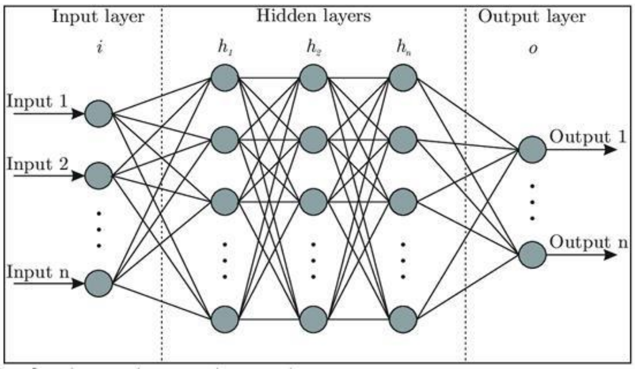


Figure 2. Artificial Neural Network Procedure

Artificial Neural Networks (ANNs) are computational models inspired by the biological neural networks of the human brain. They are highly effective in identifying and modeling non-linear and complex relationships within data,

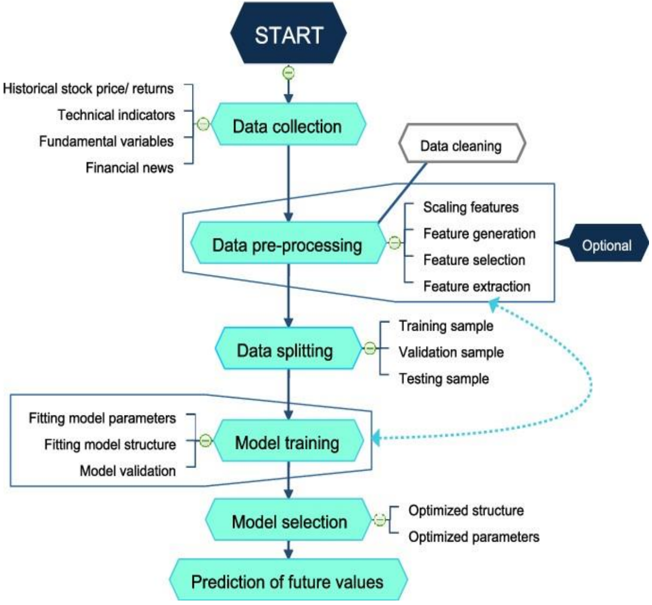


Figure 3. Flowchart for Artificial Neural Network (ANN)

making them particularly well-suited for stock price prediction. ANNs are structured into layers of interconnected neurons, where each layer processes and passes information to the next, ultimately generating predictions**.**

Input Layer:

The input layer, in this case, is referred to as the input entry point where the data is passed into the ANN. Here, every neuron represent a feature of the dataset, and these include:
Historical Stock Prices : Closing prices or high/low levels over history.
Trading Volume: The number of shares traded in a given period.
Technical Indicators: Features that are derived from financial models, such as a moving average, RSI, or Bollinger Bands.
External Factors: Inputs such as macroeconomic indicators, news sentiment, or changes in currency.The raw data is given to the network properly arranged because of the input layer.

Hidden Layers:

True nerve-like layers are hidden layers, where the core of the ANN is located, capturing meaningful patterns and relationships in the data. Every hidden layer contains multiple neurons connected to every neuron in the previous and subsequent layers.

Activation Functions:

The neurons use an activation function such as sigmoid, ReLU (Rectified Linear Unit), or tanh, to introduce the nonlinearity. These introduce non-linearity and enable the network to model stock data's more complex relationships. For instance, ReLU ensures only significant patterns propagate since all negative values are regarded as zero. On the other hand, sigmoid maps the values between 0 and 1 into the probability.

Role of Layers:

Each hidden layer does extract some kind of feature or pattern, like frequency of daily occurrence or by seasonality, which explains the generalization capacity of the network across many data conditions.

Output Layer:

The output layer is supposed to produce the final prediction for the learned patterns in the hidden layers. For a stock price, this output may include: The stock price for the following period of reference (say, next week or next month).
Probability of a positive or negative stock price (classification tasks). The output layer depends on the type of task, and contains only one neuron in the case of the regression task like the prediction of the stock price**.**

2) support vector machine

SVM is probably one of the most versatile and robust categories of supervised learning models used for classification as well as regression tasks. Their mathematical foundation and the ability to encapsulate complex relationships between data elements make them particularly effective in applications to predictive modeling in diverse domains, including even prediction of stock prices. This is the area where Support Vector Regression (SVR) is extensively used in forecasting for continuous values, like future stock prices or trends. SVR is always valuable for future predictions considering precision and robustness, especially with noisy,

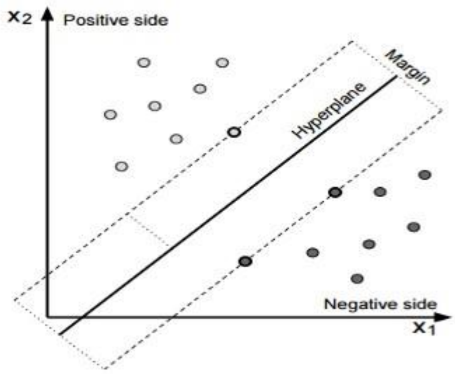


Figure 4. Architecture of Support Vector Regression highly dimensional, and non-linear financial data.pport Vector Regression (SVR) is applied to predict continuous variables, namely the stock price of the future. SVR focuses on identifying a regression function that best approximates the relationship between input features (e.g., historical prices, trading volumes, technical indicators) and target outputs (future stock values). What sets SVR apart from traditional regression models is its use of margin-based optimization to find a function that minimizes prediction error while maintaining generalization.

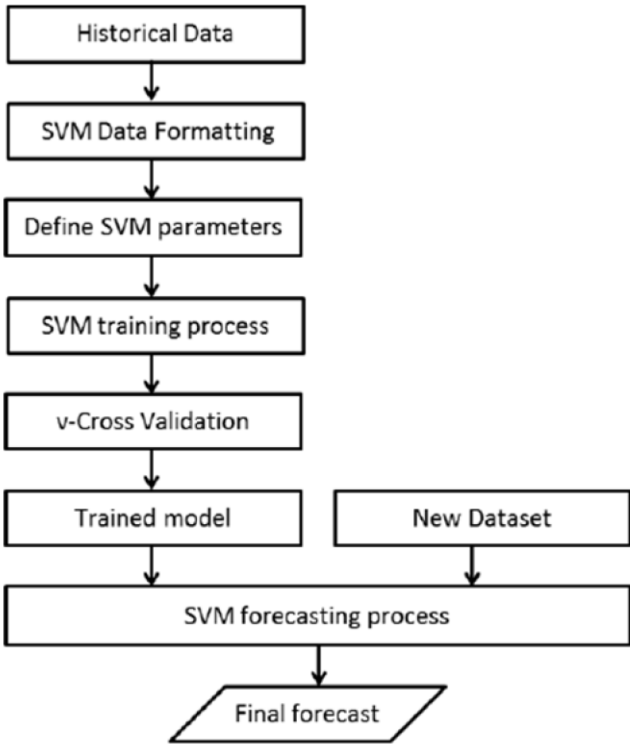


Figure 5. flowchart of Support Vector Regression

SVM is probably one of the most versatile and robust categories of supervised learning models used for classification as well as regression tasks. Their mathematical foundation and the ability to encapsulate complex relationships between data elements make them particularly effective in applications to predictive modeling in diverse domains, including even prediction of stock prices. This is the area where Support Vector Regression (SVR) is extensively used in forecasting for continuous values, like future stock prices or trends. SVR is always valuable for future predictions considering precision and robustness, especially with noisy, highly dimensional, and non-linear financial data.

It is a Supervised Machine learning algorithm used for regression analysis. It finds the function that helps us approximate mapping based on the training sample from an input domain to real numbers. The Terminologies contained in this are Hyperplane -this is the line that is used to predict the continuous output. Kernel helps to find hyperplanes in higher dimensional space without increasing the computational cost of it and the decision boundary is a simplification line that differentiates positive examples and negative examples.

1.Kernel Functions-Based High-Dimensional Transformations:

Thus, one inherent characteristic feature of SVMs is the use of kernel functions, which allows the model to transform the input data into a higher dimensional feature space. This transformation enables SVM to detect intricate, non-linear patterns within the data set that might not otherwise be evident in original data. Examples of some popular kernel functions include the following:.

Polynomial Kernel: It manages more complex relationships that include polynomial terms in the feature space, thus facilitating flexible nonlinear modeling.

Radial Basis Function (RBF) Kernel: The most used kernel in finance and time series data probably. RBF kernels are highly proficient in modeling localized, non-linear dependencies and notably perform well in high-dimensional spaces.

It is highly significant in determining whether this model captures important trends in the data of the stock prices or not. For example, financial markets hold complex, intertwined trends initiated by various factors such as macroeconomic indicators, news sentiment, and historical price fluctuations. SVMs can separate these kinds of interlinked relations with precision by mapping these features to higher dimensions.

2. Optimal Hyperplane and Regression Line:

SVMs attempt to find the best-fit hyperplane (in classification applications) or regression line for SVR, which gives the least possible prediction errors while providing as large a margin as possible. The objective of the model is to make the best decision boundary or regression function able to separate or fit these data points most appropriately.

For defining SVR, it offers an epsilon tolerance margin wherein it is not billed in case true values are violated in the range of such a margin. This ensures that the model views the general trend and tries to avoid overfitting to minor changes or noise in financial data.

This ability of definition of optimal boundary makes SVMs robust about handling financial data since the difference between noise and meaningful patterns can hardly be defined.

3) Random forest

Random forest Random Forests is a particularly robust and highly versatile type of supervised machine learning algorithm, widely applied both in classification tasks and in regression tasks. Unlike the typical single decision tree that may lead to overfitting and diversity, the ensemble approach in Random Forests aggregates the predictions of different decision trees; this ensures a much more stable and accurate result. This ensemble approach makes sure that a compounding of minimal weaknesses and the maximum strengths of individual trees is made to achieve better generalization and reliability.
Random forest is a supervised Machine Learning algorithm used in Regression. Analysis. This overcame the problem that was seen with overfitting in the decision tree [12]. It is an Ensemble learning technique. The prediction procedure goes about as follows: first, one randomly selects k data point From the training set then accordingly the decision tree is built. Then choose the number of If we wanted to build trees, then again go through the previous steps. We do from every new data point. N Tree Trees predict the value of Y for data points and assigns new data points across all ofy predicted Y values.

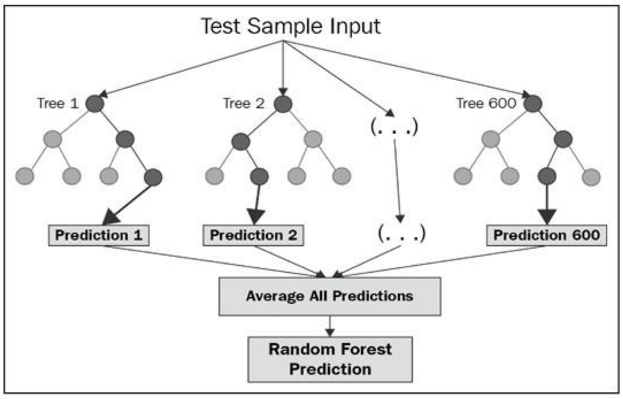
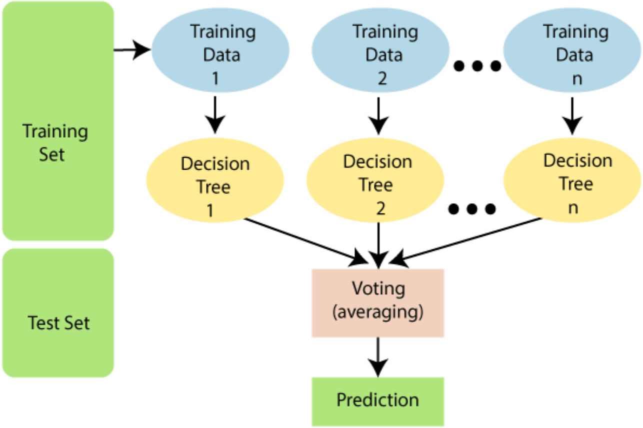


Figure 6. Random Forest Procedure

Since Random Forests intrinsically own the property of deal ing with complex and high-dimensional data, the intrinsic feature aptly makes them suitable for applying in domains likestockpriceprediction in whichthe datasets fairly depict hi gher noise and intricate relationships between variables.
The factors that influence the stock prices are historical trends, trading volumes, market sentiment, and macroeconomic indicators. Those interdependencies are then what Random Forests are pretty good at: they construct multiple decision trees, train them on data with randomized subsets and features every time, and aggregate their outputs to produce pretty-accurate predictions.



This enhances the precision of the prediction along with resistance to noise and outliers that usually dominate financial data sets. This flexibility in modeling non-linear relationships and efficiency in scaling in accordance with large datasets makes Random Forests less applicable for time-consuming tasks of challengeable predictions like the stock price and others.
RandomForests is a strong ensemble technique that can be applied to classification as well as regression tasks. The algorithm was designed specifically to combine the outputs of predictions from more than one decision tree so that it may enhance stability, accuracy, and generalization. The functionality of Random Forests is based on obtaining multiboosted decision trees by aggregating their predictions. The process takes place in the following way:

Building Decision Trees with Random Subsets:

Random Forests use a type of bootstrap aggregation called bagging to generate different decision trees. Since in a random subset data points from the training set are picked with replacement, every tree sees a different combination of data.

It additionally randomly chooses a subset of feature in each subset to consider at each split within the tree. This will tend not to have one single feature dominating the model and ensures diversity among the trees.

Tree Construction

Each tree is constructed in a recursive fashion, splitting the data according to values of preferred features so that the prediction errors are minimized (e.g., the Mean Squared Error in regression tasks). These splits are performed until a stopping criterion is reached: either the maximal depth of the tree is reached, or some minimum number of samples in a leaf node is achieved. **Aggregation of Results:**

Once the trees have been constructed, their predictions get combined to produce a final output: For Regression: The prediction of all the trees are averaged, so that the ensemble captures the general trend of data while minimizing individual tree errors. Output classification is by majority voting: The most predicted class across all trees in the forest is picked as the representative. In the Random Forests, the predictions of a number of trees are combined thereby reducing individual variance and bias of trees to give more reliable and accurate predictions overall

RESULTS AND DISCUSSION:

A. Artificial Neural Networks (ANN):

ANN showed excellent performance and was able to fit complex nonlinear patterns in the stock data and attained up to 95.70% accuracy for large multi-feature datasets.

Its architecture is effective at modeling complex interactions but sensitive to the quality of data used, and computationally expensive. ANN is excellent if used for long term forecasting where the market is stable, but it suffers with noisy data and may be less suitable for real-time prediction purposes.

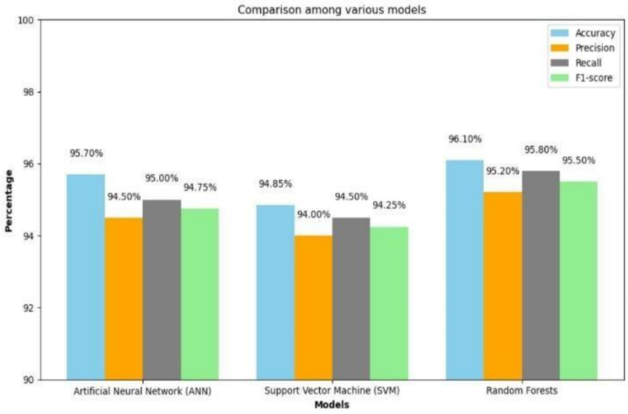
B. Support Vector Machine (SVM):

SVM achieved around 94.85% accuracy in predicting directional trends, making it effective for classifying stock price movements. Its margin-maximizing approach works well with smaller, simpler datasets, though it requires careful hyperparameter tuning. SVM's limitations with multi-dimensional data reduce its accuracy for precise price forecasts, but it remains a practical choice for short-term trend prediction with limited features.

C. Random Forest :

Random Forest reached approximately 96.10%accuracy in trend prediction and handled noisy data well due to its ensemble design. The model's reliance on multiple decision trees limits overfitting, providing stable and generalizable predictions. While it may not capture complex patterns as well as ANN, Random Forest's ease of use and reliability make it ideal for general-purpose stock predictions in fluctuating markets Random Forests are a powerful ensemble learning technique that excels in both classification and regression tasks. The algorithm is designed to combine the predictive outputs of multiple decision trees, thereby improving the stability, accuracy, and generalization of predictions. This ensemble approach makes Random Forests particularly effective in scenarios where data is noisy, complex, and influenced by numerous interacting variables, such as stock price prediction.

Model	**Accurac y**	**Precisio n**	**Recall**	**F1score**
Artificia l Neural Network	95.70%	94.50%	95.00 %	94.75 %
Support Vector Machine	94.85%	94.00%	94.50 %	94.25 %
Random Forests	96.10%	95.20%	95.80 %	95.50 %



Conclusion:

The comparative analysis of Artificial Neural Networks
(ANN), Support Vector Machines (SVM), and Random Forests in stock price prediction reveals distinct strengths and limitations for each model. ANN excels at capturing complex, non-linear patterns and is highly accurate with large, stable datasets, though it is resource-intensive and sensitive to data quality. SVM is practical for short-term trend classification in simpler datasets but requires careful tuning and struggles with high-dimensional data. Random Forest provides stable, generalizable predictions and handles noisy data effectively, though its ensemble approach limits its precision in modeling intricate patterns. Together, these insights suggest that a combined approach, leveraging each model's strengths according to specific market conditions, could enhance overall predictive accuracy and robustness in stock forecasting applications**.**

REFERENCES :

1. Subasi, A., Amir, F., Bagedo, K., Shams, A., & Sarirete, A. (2021). Stock market prediction using machine learning. Procedia Computer Science, 194, 173-179.
2. Strader, T. J., Rozycki, J. J., Root, T. H., & Huang, Y. H. J. (2020). Machine learning stock market prediction studies: review and research directions. Journal of International Technology and Information Management, 28(4), 63-83.
3. Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia computer science, 167, 599-606.
4. Kompella, S., & Chakravarthy Chilukuri, K. C. C. (2020). Stock market prediction using machine learning methods. International Journal o6f Computer Engineering and Technology, 10(3), 2019.
5. Li, Y., & Pan, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. International Journal of Data Science and Analytics, 13(2), 139-149.
6. Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53(4), 3007-3057
7. Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. Ieee Access, 8, 150199150212.
8. Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. Ieee Access, 8, 150199150212.
9. Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, 197, 116659.
10. Li, A. W., & Bastos, G. S. (2020). Stock market forecasting using deep learning and technical analysis: a systematic review. IEEE access, 8, 185232-185242.
11. Zhang, K., Zhong, G., Dong, J., Wang, S., & Wang, Y. (2019). Stock market prediction based on generative adversarial network. Procedia computer science, 147, 400-406.
12. Mehtab, S., & Sen, J. (2022). Stock price prediction using machine learning and deep learning algorithms and models. Machine Learning in the Analysis and Forecasting of Financial Time Series,
235-303
13. Rezaei, H., Faaljou, H., & Mansourfar, G. (2021). Stock price prediction using deep learning and frequency decomposition. Expert Systems with Applications, 169, 114332.
14. Bhattacharjee, I., & Bhattacharja, P. (2019, December). Stock price prediction: a comparative study between traditional statistical approach and machine learning approach. In 2019 4th international conference on electrical information and communication technology (EICT) (pp. 1-6).
IEEE.
15. Asad, M. (2015, October). Optimized stock market prediction using ensemble learning. In 2015 9Th international conference on application of information and communication technologies (AICT) (pp. 263-268). IEEE.