A study on different approaches to the Predictive analysis of stock market Using machine learning

First A. Soumen Maity,

B. Tech CSE, JIS UNIVERSITY,

soumenm.cse.jisu21@gmail.com

Third C. Ujjwal Dinda

B.Tech CSE, JIS UNIVERSITY,

ujjwald.cse.jisu21@gmail.com

Fifth E. Ananya Dhar

B.Tech CSE, JIS UNIVERSITY,

ananyad.cse.jisu21@gmail.com

Second B. Sounak Dey,

B.Tech CSE, JIS UNIVERSITY,

sounakd.cse.jisu21@gmail.com

Fourth D. Amitava Singha

B.Tech CSE, JIS UNIVERSITY,

amitavas.cse.jisu21@gmail.com

Sixth F. Dr. Paramita Sarkar

Asst. Prof. Of JIS UNIVERSITY,

mailtoparo@gmail.com

***Abstract*— As we move into the future many high technological innovations and analysis formulas are taking place in stock market analysis and transforms into a new era of stock trading. As many cross border open market policies are evolving making the market growth fast and attractive for more investors. That's why many common people and investors are considering stock market as a smart financial investment plan for future. But investing in stocks required many financial analysis and time to eliminate this problem and make stock investment more easy, many researchers and programmers are building many online tools, formulas and analysis patterns. However, adding machine learning algorithms for smart and big analysis is considered to be the most effective and modern approach. This Technique consists of many models that allows to work with non-traditional data, large data based on stocks performance. Advanced ML techniques offers more features like text to data transform, NLP or social media data extraction, sentiment analysis, Gio Political conditions and many more. But Still stock market forecasting and analysis is a complex region as it's a ongoing process consists of large, unstable, incomplete data. This paper explores machine learning-based approaches for stock market prediction through the implementation of a generic framework. Research findings from the few past decades historical data were critically examined, sourced from digital libraries and databases like ACM Digital Library and Scopus. Additionally, we conducted some model's performance test and similarities with other models and their forecasted results.**

Keywords—stock market forecasting, ARIMA model , GARCH model, LSTM, CNN model, Random Forest.

I. INTRODUCTION

From the past few decades information technology had made a huge advancement in businesses processes, Global financial market is became one of the most influential developed sector in recent time. In recent years, investing in stocks has gained Some attention of individual investors largely due to technological innovations. Investors are increasingly seeking tools and techniques to maximize profits and minimize risks. Though stock market prediction in still illegal and unpredictable as its movement is totally volatile and chaotic. There are infinite numbers of constraints that can affect the stock movement. Here we are discussing about stock market prediction based on historical price data. As this price information covers a big portion of stock nature, that's why incorporating analysis with this information leads to a forecasting possibility. This analysis helps in getting more accurate prediction which is most important for driving profits. It is clear that there is a conflict between acceptance of historic data as analysis information. Some researchers believed that stocks follows random walks according to efficient market hypothesis. But most researchers also proved that historic stock data can be acceptable to very extent and make this best fit for analysis. So LSTM models are getting an advantage over traditional analysis.

Although this theory was widely accepted in the past, technological advancements have shown that stock prices can be predicted to some extent. By analyzing historical market data alongside information from social media platforms, researchers have been able to forecast changes in economic and business sectors. The success of stock market prediction systems heavily depends on the quality of the features they use. While some methods have been employed to enhance stock specific features, more focus is needed on improving feature extraction and selection processes.

II. EASE OF USE

The Stock Market prediction historical data has been composed from yahooh finance datasets. 10 years of financial data are formed, preprocessed using existing techniques, and then normalized, resized, and histogram equalization techniques. After finishing the data compilation, we apply prediction models means, ARIMA model, ANN, GARCH that have been useful in finding the prediction Maintaining the Integrity.

"Xu, C.; Li, J.; Feng, B.; Lu, B. A Financial Time-Series Prediction Model Based on Multiplex Attention and Linear Transformer Structure" [1] proposed a linear model where Linear transformers reduce the time complexity of the model, allowing faster training and inference, its multiplexed attention mechanism improves the model's ability to capture long-term dependencies, resulting in higher prediction accuracy. The model handles large datasets efficiently, making it applicable to real-world financial forecasting. [1]

"Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar" [2] highlighted that the ANN was shown to provide better prediction accuracy compared to RF, especially for companies like Nike, JP Morgan, and Pfizer. Creating new variables (moving averages and price deviations) helped improve the models' predictive performance. The use of RMSE, MAPE, and MBE allowed for precise evaluation of the prediction errors.[2]

"A. A. Ariyo, A. O. Adewumi and C. K. Ayo "[3] suggested an approach that ARIMA models are particularly effective for short-term stock price prediction. The methodology of model identification, parameter tuning, and residual analysis is critical to developing a successful ARIMA model. For both Nokia and Zenith Bank stock indices, ARIMA models provided accurate short-term forecasts, aiding investment decision-making.

"Md. Mobin Akhtar, Abu Sarwar Zamani, Shakir Khan, Abdallah Saleh Ali Shatat, Sara Dilshad, Faizan Samdani" [4] proposed system of machine learning algorithms using statistical data of stock market which uses Random Forest and SVM algorithms to predict stock prices based on historical data. The data is pre-processed to remove inconsistencies, and relevant features such as opening price, closing price, volume, and high/low values are used to train the models. The system architecture is built to handle large datasets, sourced from Kaggle, and prepares them for training and testing. The models aim to predict the stock price for particular time and evaluate the performance using accuracy metrics.[4]

"Chung, H., Shin, Ks. [5] mainly focuses on Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction using multi-channel CNN network models", where Genetic Algorithms effectively optimized CNN architectures, leading to improved accuracy in stock market predictions. Multi-channel CNNs were demonstrated to be highly efficient in capturing local temporal patterns in stock data. The proposed method outperformed both ANN and standard CNN models, proving the strength of the hybrid GA-CNN approach. [5]

"Mukherjee, S., Sadhukhan, B., Sarkar, N., Roy, D., & De, S." [6] enlightened on Stock market prediction using deep learning algorithms that CNN model's ability to retain time-series context through its convolutional layers allowed it to outperform the ANN model, which tends to lose context. The 2-D histogram approach used in the CNN provided more accurate predictions by capturing complex patterns in the stock data. The study found that the CNN model was also less prone to overfitting, thanks to the use of regularization techniques. The CNN model also demonstrated faster training and better performance in extreme market conditions, such as during the COVID-19 pandemic. The case study revealed that although the CNN model's accuracy dropped to around 91% during the initial days of the market crash, it quickly adapted and achieved 95%–98% accuracy as the market stabilized.[6]

A. *Comparison Table*

Paper Ref No.	Key Attributes	Results
Appl. Sci. 2023, 13, 5175. [1]	GARCH model volatility. Multilayer perception (MLP) and Recurrent neural networks (RNN). [1]	Computational Efficiency, Prediction Accuracy, Inference Speed.
"Volume 167, 2020, Pages 599- 606, ISSN 1877- 0509." [2]	Random Forest Artificial neural network	RMSE value for ANN is 0.42, and MAPE is 0.77%
"2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2014, pp. 106-112." [3]	Autoregressive component and ARIMA model.	High effectiveness of ARIMA model for short-time data.
"Science, Volume 34, Issue 4, 2022, 101940, ISSN 1018-3647." [4]	Input information is stock historic data and used frameworks are SVM and RF.	As the result we got effectiveness of 80.8% from RF and 78.7% form SVM.
"Neural Comput & Applic 32, 7897–7914 (2020)." [5]	CNN Architecture and Genetic Algorithm for Optimization.	GA-CNN 73.74%, CNN 70.16% and for ANN 58.62%.
"CAAI Transactions on Intelligence Technology, 8(1), 82-94." [6]	ANN with Backpropagation and CNN with 2-D Histograms.	ANN Accuracy: 97.66%. CNN Accuracy: 98.92%.

Table 1: Comparison Table

The comparison between stock price prediction models using ARIMA, ANN, and Random Forest (RF), GARCH model, CNN techniques highlight the strengths and limitations of each method, particularly for short-term forecasting.

III. Model Description

We developed the model in various steps using historical datasets of stock prices. First, we take the historical data then start to structure it in a way that no null value will be present. Now we slice the datasets in smaller chunks and start plotting them in models. After segmentation, feature extraction using time-series analysis and auto ARIMA for price prediction.

A. *Datasets*

Yahoo finance offers a wide range of historical data of all stocks and indexes in any desired range. It supports some useful features like offline download of historic data in csv format for future analysis. It also supports API connection for real- time historical stocks, forex, and cryptocurrency data. This datasets are consists of open, close, high, low price in a particular day. Also has dividends and split data functionality. Some other famous websites are alpha vantage, quand1, google finance.

date	Open	high	low	close	Adj close	volume
2019-9-3	10960.95	10967.50	10772.70	10797.90	10797.90	483000
2019-9-4	10790.40	10858.75	10746.34	10844.65	10844.65	508800
2019-9-5	10860.95	10920.09	10816.00	10847.90	10847.90	595700

……	……	……	….. .	……	……	…….
2024-8-28	25030.80	25129.59	24964.65	25052.34	25052.34	220400
2024-8-29	25035.30	25192.90	24998.50	25151.94	25151.94	354000
2024-8-30	25249.69	25268.34	25199.40	25235.90	25235.90	638200

Table 2:- Nifty50 Stock price table

Now we get our data set and import all the necessary modules and extensions. Then the very first step is to plot the data into a graph.

Moving to our next step that is test for stationarity as our selected ARIMA model works on non-stationary data, for further the researchers divided this process into three parts. As follows level (the average value in the series), trend (the increasing or falling value in the series) and seasonality (recurring short-term cycle). In addition many researchers are also count noise (the random variance in the series) as a component. In case of non-stationary data eliminate the null data and other noise so that at time of model implication no error generates. Generating error leads to poor prediction.

we perform a stationarity test with the Augmented Dickey- Fuller (ADF) test, which helps us confirm if the data series is stationary. If non-stationary, adjustments like differencing and logarithmic transformation are applied to ensure stability, making the data fit for ARIMA modeling.

After confirming the need to remove trends and seasonality, we decomposed the time series into these components. This enabled us to isolate and observe the underlying patterns, providing insight into the seasonality and trends that influence stock prices. To further prepare the data, we applied a log transformation to reduce the range and stabilize the trend. Using rolling averages, we identified the mean and standard deviation over time, further validating the transformation's effectiveness.

We then split the data into training and testing sets to evaluate the model's accuracy. We used the auto_arima function to automate the parameter selection for the ARIMA model, which efficiently finds the best p, d, and q values (lag observations, differencing order, and moving average order, respectively). The model was optimized with the values 1, 1, and 2 for p, d, and q, respectively, based on the best-fit results of this function. This automated process enhances precision by eliminating manual parameter adjustments, making it easier to find the optimal configuration for accurate predictions.

In our experiment we use both index and stock values. Index values lead to more noise as it comes with more volatility in comparison of stock values. At index value we use nifty50 price for the period of September 2019 to august 2024 (1235 rows), which gives the model around 5years of historical data to work with. We separated all the values changes in a day like high, low, open, close and volume to give the proper insight of market movement.

For minimizing the investment risk and informed trading strategy ARIMA model framework excels and supports analysis of little market movements.

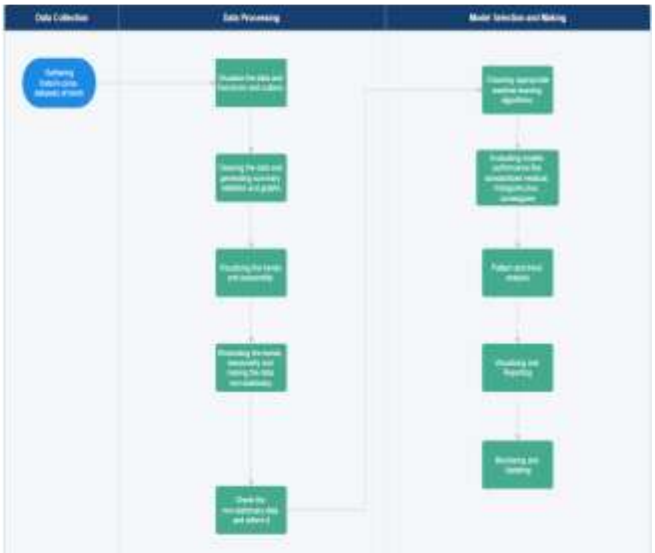

Fig 1: - Flowchart of the machine learning model.

IV. Result Analysis

The ARIMA model, which stands for Autoregressive Integrated Moving Average, is a powerful tool for forecasting stock prices based on historical data, especially for short-term investment decisions. we applied ARIMA to predict the stock prices of nifty50 by breaking down the model's structure and methodology into three key components: autoregression (AR), differencing (I), and moving average (MA). The AR component analyzes the relationship between the current and past stock prices, helping to recognize recurring patterns. Differencing helps make the time series stationary by reducing the effects of trends and seasonality. Finally, the MA component addresses noise by accounting for past forecast errors, which smoothens the prediction output. Together, these elements make ARIMA adept at capturing dynamic trends in stock price movements, supporting accurate predictions of future price levels.

The first step involved visualizing the stock's daily closing prices. Plotting these prices revealed an upward trend with fluctuations over time, indicating a non-stationary time series. Because time series forecasting requires stationary data, we conducted the Augmented Dickey-Fuller (ADF) test to test for stationarity. The ADF test checks for a "unit root," which indicates whether the data shows systematic trends or seasonal cycles. Our results, showing a high p- value and a test statistic exceeding critical values, confirmed non-stationarity. This led us to perform differencing, removing trends and stabilizing variance, which allows the ARIMA model to perform reliably.

Four initial models are fitted: ARIMA (0, d,0), ARIMA (2, d,2), ARIMA (1, d,0), ARIMA (0, d,1), A constant is included unless d=2. If d≤1, an additional model is also fitted: ARIMA (0, d,0) without a constant. The best model (with the smallest AICc value) fitted is set to be the ―current model‖.

A non-seasonal ARIMA model can be written as

$$y_t = \theta_0 + \varnothing_1 y_{t-2} + \cdots + \varnothing_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

The autoregressive operator is defined as:

$$\varnothing_B = 1 - \varnothing_1 B - \varnothing_2 B^2 - \cdots \varnothing_P B^p = 1 - \sum_{j=1} \varnothing_j B^j \quad (2)$$

or equivalently as :- $\varnothing(t) X_t = w_t$

where $c=\mu(1-\phi_1-\cdots-\phi_p)$ and μ is the mean of $(1-B)dy_t$. R uses the parameterization of Equation.

We can re-write Equation as

$$\phi(B)(1-B)dy_t=c+\theta(B)\varepsilon_t \quad (3)$$

where $\phi(B)=(1-\phi_1 B-\cdots-\phi_p B^p)$ is a p^{th} order polynomial in B and $\theta(B)=(1+\theta_1 B+\ldots+\theta_q B^q)$ is a q^{th} order polynomial in B.

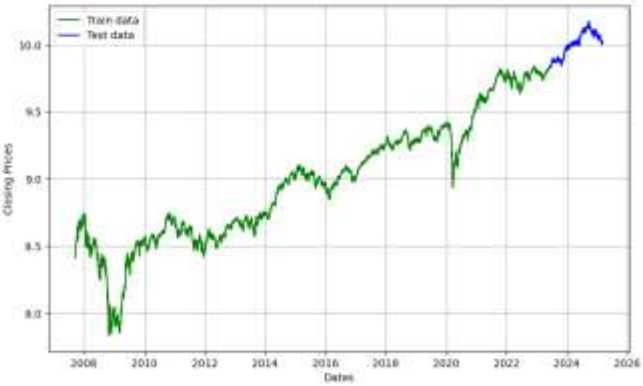

Fig 2:- Train data and test data

After training the model on the training set, we made forecasts for the testing set [as in fig 2 and fig 3] with a 95% confidence interval, plotting predicted prices against actual values. This visualized forecast helped us identify areas where the model closely mirrored actual stock prices, showcasing its predictive accuracy. Finally, we evaluated model performance with common metrics like mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). The low MAPE of 2.5% indicated the model's 97.5% accuracy in forecasting stock prices, confirming its robustness for time-series analysis in stock forecasting.

model	accuracy
Random Forest (RF)	85%
GASVM	93.7%
ANNG	30%
ANN, SVM, LSTM	65%
MLP	64%
ARIMA	97.5%
DL	86%

Table 3: - Comparison table of different models results

performance	value
MSE	0.015076627702326598
MAE	0.1150099911420342
RMSE	0.12278691991546412
MAPE	0.025397463962705633

Table 4: - Performance table of our model

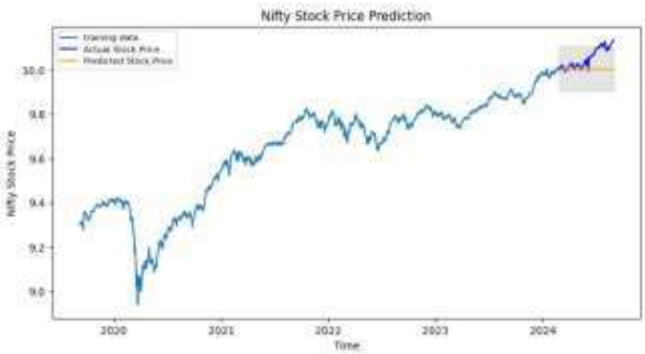

Fig 3: - Model forecasting

V. CONCLUSION AND FUTURE WORK

This paper is primarily focused on the various LSTM frameworks like - ARIMA model, GARCH model, RF, ADF, ANN, CNN and their performance. As discussed earlier LSTM supports efficient forecasting techniques which helps in further analysis. As we address the main issue regarding decision making on investment this models helps investors to understand in where and when they should invest. The result like forecasted graph, sarimax results and other analysis values shows that ARIMA model is best suited for this type of short time predictions.

For the future work hybrid models (ANN+ARIMA or ARIMA+RF) can improve the prediction accuracy and efficiency. As stock price is also aligned with human sentiment and global financial crisis so hybrid frameworks with NLP and sentiment analysis or social media data can boost the prediction accuracy in a remarkable way.

REFERENCES

[1] Xu, C.; Li, J.; Feng, B.; Lu, B. A Financial Time-Series Prediction Model Based on Multiplex Attention and Linear Transformer Structure. Appl. Sci. 2023, 13, 5175. https://doi.org/10.3390/app13085175

[2] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, Stock Closing Price Prediction using Machine Learning Techniques, Procedia Computer Science, Volume 167, 2020, Pages 599-606, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.326. (https://www.sciencedirect.com/science/article/pii/S1877050920307924)

[3] A. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge, UK, 2014,pp. 106-112, https://doi.org/10.1109/UKSim.2014.67.keywords:{Predictivemodels; Indexes;Timeseries analysis;Forecasting;Data models;Computationalmodelling;Mathematical model;ARIMA model;Stock Price prediction;Stockmarket;Short-term prediction}

[4] Md. Mobin Akhtar, Abu Sarwar Zamani, Shakir Khan, Abdallah Saleh AliShatat, Sara Dilshad, Faizan Samdani, Stock market prediction based onstatistical data using machine learning algorithms, Journal of King SaudUniversity - Science, Volume 34, Issue 4, 2022, 101940, ISSN 10183647, https://doi.org/10.1016/j.jksus.2022.101940.(https://www.sciencedirect.com/science/article/pii/S1018364722001215)

[5] Chung, H., Shin, Ks. Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction. Neural Comput & Applic 32, 7897–7914 (2020). https://doi.org/10.1007/s00521-019-04236-3

[6] Mukherjee, S., Sadhukhan, B., Sarkar, N., Roy, D., & De, S. (2021). Stock market prediction using deep learning algorithms. CAAI Transactions onIntelligence Technology, 8(1), 82-94. https://doi.org/10.1049/cit2.12059

[7] Khan AH, Shah A, Ali A, Shahid R, Zahid ZU, Sharif MU, Jan T, Zafar MH. A performance comparison of machine learning models for stock market prediction with novel investment strategy. PLoS One. 2023 Sep 21;18(9):e0286362. doi: 10.1371/journal.pone.0286362. PMID: 37733720; PMCID: PMC10513304.

[8] Xiao, Daiyou, Su, Jinxia, Research on Stock Price Time Series Prediction Based on Deep Learning and Autoregressive Integrated Moving Average, Scientific Programming, 2022, 4758698, 12 pages, 2022. https://doi.org/10.1155/2022/4758698

[9] Li, Q., Kamaruddin, N., Yuhaniz, S.S. et al. Forecasting stock prices changes using long-short term memory neural network with symbolic genetic programming. Sci Rep https://doi.org/10.1038/s41598-023-50783-0

[10] STOCK MARKET ANALYSIS THROUGH MACHINE LEARNING ALGORITHMS http://ir.juit.ac.in:8080/jspui/jspui/handle/123456789/7730

[11] Malti Bansal, Apoorva Goyal, Apoorva Choudhary, Stock Market Prediction with High Accuracy using Machine Learning Techniques, Procedia Computer Science, Volume 215, 2022, Pages 247-265, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2022.12.028. (https://www.sciencedirect.com/science/article/pii/S1877050922020993)

[12] Phuoc, T., Anh, P.T.K., Tam, P.H. et al. Applying machine learning algorithms to predict the stock price trend in the stock market – The

case of Vietnam. Humanit Soc Sci Commun https://doi.org/10.1057/s41599-024-02807-x

[13] Mroua, M., Lamine, A. Financial time series prediction under Covid-19 pandemic crisis with Long Short-Term Memory (LSTM) network. Humanit Soc Sci Commun 10, 530 (2023). https://doi.org/10.1057/s41599-023-02042-w

[14] Chopra, R.; Sharma, G.D. Application of Artificial Intelligence in Stock Market Forecasting: A Critique, Review, and Research Agenda. J. Risk Financial Manag. 2021, 14, 526. https://doi.org/10.3390/jrfm14110526