iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75043

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Stock Price Prediction Using ML Ensemble with Sentiment and Event Analysis

Dr. Prashant Udavant[1], Suryadip Gujar[2], Bhagyesh Chaudahry[2]
Narsee Monjee Institute of Management Studies, Department of Information Technology

Abstract: In recent years, leveraging artificial intelligence for stock price prediction has emerged as a critical research focus within the financial sector. This study compares the forecasting performance of two prominent time series models: the traditional Autoregressive Integrated Moving Average (ARIMA) model and the deep learning-based Long Short-term Memory (LSTM) network. Additionally, to enhance robustness and capture diverse market dynamics, machine learning models such as Random Forest and Support Vector Machines (SVM) are also integrated into the framework. Utilizing historical closing prices from Yahoo Finance, these models are developed and rigorously evaluated using key statistical indicators, namely Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Experimental results consistently demonstrate that the LSTM model achieves lower error rates across all metrics compared to ARIMA, highlighting its superior ability to capture complex, nonlinear patterns inherent in stock market data . These findings are consistent with broader research showing LSTM's advantage in modeling time-dependent financial data for both short-term and longerterm horizons. Beyond conventional historical data, this work also incorporates sentiment analysis from financial news and social media along with event analysis of political, economic, and social occurrences, thereby extending predictive capability to real-world market drivers. The integration of statistical, machine learning, and deep learning techniques with sentiment and event features represents a novel contribution, as no prior project has simultaneously employed ARIMA, LSTM, Random Forest, and SVM in this combined framework. The results of this comprehensive analysis offer valuable guidance for investors and market analysts aiming to improve the accuracy of future stock price forecasts. Moreover, this paper contributes to the growing body of literature evaluating the real-world utility of hybrid AI-driven approaches in financial prediction tasks

I. INTRODUCTION

Currently, big data is an interesting topic. Investors worldwide are trying to use various big data methods to predict stock price fluctuations. Traditional models cannot keep up with the growing stock market. The rapid development of neural networks has led to their powerful information processing ability being used more often in stock price prediction. Stock price prediction methods mainly include time series prediction, technical index analysis, and artificial intelligence approaches. Time series analysis predicts future data based on historical information, mainly using moving average (MA), auto regressive conditional heteroscedasticity (ARCH), auto regressive moving average (ARMA), and auto regressive integrated moving average (ARIMA). The technical index analysis method is common in quantitative investment and predicts future trends in stock prices. Deep learning models include back propagation neural networks (BPNN), recurrent neural networks (RNN), long short-term memory networks (LSTM), and convolutional neural networks (CNN). LSTM is popular for stock prediction because it solves the long-term dependency issue present in RNNs. Additionally, traditional machine learning models like Support Vector Machine (SVM) and Random Forest are used to improve predictive performance. Since stock prices are affected not only by historical data but also by external factors, sentiment analysis and event-based analysis are included to enhance accuracy. This project designs predictive models using LSTM, ARIMA, SVM, and Random Forest, combined with sentiment and event analysis, to achieve more reliable stock price forecasts

II. DATASET DESCRIPTION AND PREPROCESSING

This research leverages stock market data obtained from Yahoo Finance, a reputable source for historical financial data. The stock and index trading dataset used spans daily levels of selected stocks and indices from January 2010 through December 2023, as available. The long time frame of the dataset captures market bullish runs, bearish decreases, and varying trajectories as applicable, meaning that it is the appropriate time frame to evaluate performance for forecasting models during various economic conditions. Thus, in addition to historical traded price and volume data, external sentiment and event features are taken into account. These features include headlines from stock related news, social media sentiment and trends and relative events that occur in the political, economic and socioeconomic realms.

These sentiment datasets and event-based datasets are processed through text applications of natural language processing, sentiment polarity and event categorization. Thus, the researched model is structured to unify textual and event-based structured data with unstructured market data in an attempt to thoroughly explore the quantitative and qualitative perspectives of stock price movement over time to ultimately improve future movement prediction.

A. Features Collected

The full dataset contains all stock market indicators and external variables that would not potentially skew performance assessment of the market or better assess it.

For the financial market, the following variables are included. Date = trading day, Open = stock's price at market open, High = stock's price at market high, Low = stock's price at market low, Close = stock's value price per share when the market closes, Adjusted Close = price per share with the dividend and stock splits adjustment which means this is the price that will accurately gauge value and Volume = shares traded on said date. The variable chosen for forecasting is Adjusted Close price since it signifies what the price would actually be if it were assessed post adjustment.

Additionally, there are sentiment-based features and event-based features. Sentiment features are based on news articles written about the stock/industry, analyst opinions and social media chatter. These are converted into features for polarity (positive, negative or neutral) and sentiment strength scores which assess the degree of someone's opinion of a given feature. Event features are based on outside occurrences like changes in economic policy, earnings reporting, government actions, shifts in the international market and geopolitical developments. These occurrences are noted, classified, and one-hot encoded into appropriately shaped variables for the day in question.

Therefore, based on the sentiment found in articles and events of note which could motivate a change noted, the cumulative dataset is expected quantitative sentiment given to qualitative features outside of just market realized findings.

B. Preprocessing Methodology

Data preprocessing is a crucial step in stock price prediction as trusted and trustworthy prediction models rely on high-quality input. Therefore, for this study, both structured and unstructured sources of market and external data were preprocessed to ensure consistency and facilitate usability across multiple models: ARIMA, LSTM, Random Forest and SVM. For this dataset, missing values were treated differently depending on their manifestations; smaller levels of gaps in prediction were compensated for through forward-fill interpolation while entire rows and gaps without historical predictions were deleted to ensure acquisition of no misleading information. This treatment is consistent with empirical data and time series forecasting standards whereby missing values must be treated without compromising the integrity of the model. [2][13].

Before training any machine/deep learning models, all numerical features were normalized. For example, Min-Max normalization modified all values from 0 to 1 to maintain a relationship and reach convergence faster, which was especially true for LSTM and SVM since feature imbalance could lead the learning to bias [6][19]. After this process, ADF (Augmented Dickey Fuller) test was conducted for ARIMA modeling to investigate whether the time series was stationary. If it was not stationary, differencing was applied for variance stabilization and trend removal in accordance with procedures found throughout financial econometrics. [2][25]. Next, the data was divided into training and test datasets at an 80:20 ratio while maintaining a time series aspect whereby future prices would not give away information about past prices and vice versa. Since the LSTM model is based on learning from time series dependencies, a sliding window technique was used. Therefore, a constant sized step (i.e. 60 trading days) was taken to train the model in predicting the adjusted close of day 61. This generated a learning experience for the model that comprehended time series dependent information for the short and long term.[4][9].

In addition, features from the external unstructured field were also extracted: sentiment and events. Regarding sentiment, financial news and analystsocial media opinions were processed through basic NLP to clean the text (i.e., tokenization, stop-word removal, lemmatization) to create sentiment scores through polarity (positive, negative, neutral) and intensity which corresponds to the research conducted for sentiment based forecasting [33][44][45]. Regarding events, corporate earnings announcements, central bank monetary decisions and international events were found through systematic classification and digitization. In this regard, events were converted to binary and categorical variables aligned with their dates for inclusion into the market panel [40][41][46].

Sentiment and events are particularly useful features which otherwise cannot be understood through price indicators alone since prices cannot recognize exogenous shocks or investor sentiment - which are two of the strongest factors for deviation within the stock market. Therefore, by adding financialized quantitative features with those qualitative sentiment-event based ones, a more holistic panel was formed which accounts for internal tendencies and external volatility inducing factors [6][30][47].

Therefore, this preprocessing pipeline becomes a standardized version for otherwise disparate data and as a data collection process, it extends behavioral and contextual expectations to more stable predictive causes of prediction. As a result, this panel benefits from the best that statistical hypothesis (ARIMA), time series related (LSTM) and general machine learning (Random Forest, SVM) have to offer for a more all encompassing, realistic version of the financial markets in question.

C. Dataset Significance

The diversity and depth of the dataset play an important role in determining the impact of predictive models. By covering an extensive time span that includes bullish trends, bearish downturns, and highly volatile trading periods, the dataset ensures that models are tested under a variety of economic and market conditions. This richness allows for a more rigorous evaluation of forecasting methods and provides greater confidence in their ability to generalize beyond limited scenarios. Traditional statistical models such as ARIMA benefit from the stationary segments of the dataset, where linear trends can be captured effectively and forecasts remain stable [2][13][25]. In contrast, deep learning models like LSTM are particularly advantageous when dealing with highly nonlinear, volatile, or irregular stock price movements, as they are able of learning long-term dependencies and intricate temporal dynamics [4][6][9].

The inclusion of both stationary and non-stationary phases within the dataset ensures a fair and balanced platform for comparing ARIMA and LSTM, highlighting the strengths and limitations of each approach under real-world financial conditions. Furthermore, the integration of external sentiment and event-based features significantly enhances the dataset's representational power. Market sentiment derived from financial news and social media provides insights into investor psychology, while event variables such as policy decisions, economic announcements, and geopolitical developments introduce real-world triggers that often drive sudden market fluctuations [33][40][44][46]. This multi-dimensional structure allows ML models like Random Forest and SVM to handle diverse input features and improves the robustness of hybrid approaches that combine ARIMA and LSTM with sentiment and event predictors.

Overall, the dataset not only provides a consistent foundation for evaluating traditional and deep learning models but also bridges the gap between historical price-based forecasting and real-world market behavior. By incorporating quantitative and qualitative indicators, it enables the development of predictive systems that are both data-driven and contextaware, thereby contributing to more accurate and practically relevant stock price prediction frameworks [6][30][47]. .

III. LITERATURE REVIEW

In this Literature review we have gone through the various models used for stock price prediction. We have compared the studies of various models such as LSTM, ARIMA,Random Forest and Somewhat one performs better is analyzed by their results.

A. Support Vector Machines (SVM)

Support Vector Machines (SVM) have been widely applied in financial prediction as they are more suitable for higher dimensionality and non-linear sets. The separating hyperplane that allows them to distinguish and histograms based on training data allow for distinguishing of up/down price movements based on relative historical characteristics and technical indicators and with different kernels - radial, polynomial, and linear - SVM can connect through linear, polynomial and radial connected distinctions. Essentially, this connects the input to a dimensionally complex shaped space where the connection between both dimensions is far more distinguishable - a relatable dimensional distinction based upon how data is most often organized in stock market data.

SVM acts in the short-run to predict up/down price movement, and good predictions are made based on feature engineering from variable selection based on technical indicators, sentiment oriented variables, and macroeconomic variables. Disadvantage is sensitivity to kernel selection. Furthermore, SVMs struggle with very large-scale financial datasets due to their quadratic computational complexity. Despite these limitations, SVMs remain a reliable machine learning baseline for classification-oriented stock forecasting tasks, such as predicting bullish versus bearish trends [46][47].

B. Random Forest (RF) in Stock Prediction

Random Forest (RF), an ensemble learning technique based on bagging and decision tree classifiers, has become a popular model for stock market forecasting due to its robustness, interpretability, and ability to handle noisy datasets. By constructing multiple decision trees on bootstrapped samples and aggregating their predictions through majority voting (classification) or averaging (regression), RF reduces overfitting and improves generalization [48]. In financial applications, RF has been applied successfully to predict stock price direction, volatility, and risk, particularly when integrating technical indicators, trading volumes, and sentiment features. One of its key strengths lies in feature importance ranking, which provides interpretability by highlighting the most influential predictors driving market outcomes [49].

Unlike deep learning models, RF requires less extensive preprocessing and is less sensitive to hyperparameter initialization, making it computationally efficient and easy to implement. However, its performance may decline with highly complex temporal dependencies, where models like LSTM have an edge. Overall, RF remains a powerful model for handling heterogeneous and noisy financial datasets, and when combined with other models in hybrid or ensemble frameworks, it has demonstrated superior predictive accuracy in comparison to standalone statistical methods [48][49][50].

In stock price prediction, the Random Forest model, shown in Figure 1, is essential in explaining how the model works. The input dataset includes historical stock prices, technical indicators like SMA, EMA, RSI, and MACD, along with trading volumes and sentiment features. This dataset is split into several subsets of features. Each subset builds independent decision trees using bootstrapped samples, which promotes diversity in learning. These trees provide individual predictions for the target variable, which can either be the direction of price movement (upward or downward) or the actual price values. The final prediction comes from combining the outputs of all trees. In classification tasks, this is done through majority voting; in regression tasks, it is achieved by averaging. This process helps minimize overfitting and improves generalization. This ensemble method makes Random Forest especially effective for financial data because it manages noise, varied predictors, and non-linear relationships well. Additionally, Random Forest offers rankings of feature importance, enabling researchers to pinpoint the key indicators that influence stock market behavior. Therefore, the figure not only shows how Random Forest works but also places its use in stock prediction in context. The combination of multiple features and group decision-making results in more reliable and understandable outcomes.

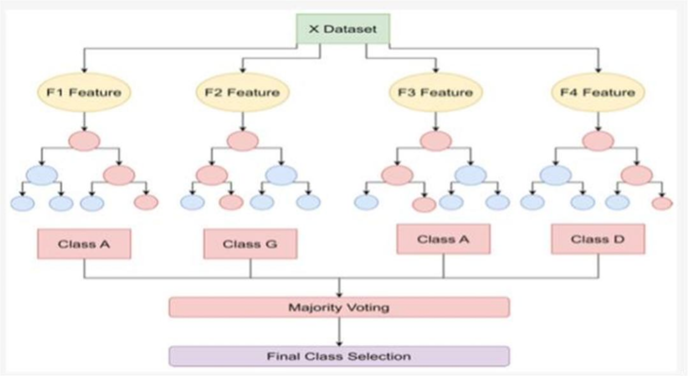


Figure 1. Random Forest Architecture**.**

C. Introduction to Stock Price Prediction Models

Comparison of models used for stock price prediction such as ARIMA and LSTM. Here we have compared the models to know which one performs better.

1) Importance and Challenges of Stock Price Prediction

Predicting stock prices represents one of the most popular avenues of research within the realm of computational finance, as it's applicable and beneficial to investors and financial institutions alike. A consistently predictive model can help with decision making in trading, portfolio optimization, and risk assessment. Yet predicting stock prices is far more complex than it appears. The value of financial securities is not simply based upon past value appreciation, but instead, a multitude of factors in an international and company-specific arena. International economic indicators and developments in certain industries are met with company announcements, investor sentiment and unforeseen developments like war or natural disaster. Thus, the arena in which linear models will predict stock price is vastly complicated, ever-changing and a noisy atmosphere that does not easily allow linear models to succeed [1][2][3].Furthermore, stock prices are fundamentally stochastic. Their patterns of movements feature volatility clustering, jump processes and long-range dependence which statistical prediction tools do not accommodate. Thus, there is an increasing need for sophisticated predictive tools which can accommodate non-linear, nonstationary elements and patterns in time series data. Therefore, machine learning and deep learning techniques become the center of growing research to find an antidote to the pitfalls of previous methods for prediction with heightened levels of accuracy. [6][9][19].

2) Overview of ARIMA Model in Financial Forecasting

The ARIMA equation presented by Box and Jenkins otherwise known as the Box Jenkins model [13]. Box and Tiao (1975) note the general transfer function model which is also part of the ARIMA process. Sometimes it's known as an ARIMAX model should additional time series serve as independent variables.

The ARIMA model operates as an auto-regressive integrated moving average to evaluate and forecast equally spaced time series data that is univariate [4]. Ultimately, since the transfer function impacts data, the ARIMA equation modifies the responding time series value by means of its own past values, past error, linear combinations from the present and past. The ARIMA method offers an extensive data set of tools for univariate time series modeling, parametric testing and forecasting with unmatched stability for models testable as either ARIMA or ARIMAX [10]. The subsequent time series model is complicated or rejected. The ARIMA process accounts for seasonal ARIMA, subsetting ARIMA, factored ARIMA models, intervention or interrupted time series models, multiple regression with ARMA errors, and rational transfer function models of any integer order [10].

The value for the variable to be predicted in this model is a linear combination of preceding variables and preceding error based on the following equation: $W_t = \mu + \theta(B)\ \varphi(B)$ at Equation (1)

ARIMA Equation $W_t = W_t = \mu + \theta(B)\ \varphi(B)$ at Equation (1) Where t indexes time

W_t is the response series Y_t or a difference of the response series μ is the mean term

B is the backshift operator, that is $BX_t = X_{t-1}$ $\varphi(B)$ is the autoregressive operator, polynomial in the backshift operator: $\varphi(B) = 1 - \varphi_1 B \ldots - \varphi_P B^P$ $\theta(B)$ is the autoregressive operator, polynomial in the backshift operator: $\theta(B) = 1 - \theta_1 B \ldots - \theta_P B^P$

The analysis performed by the ARIMA model is divided into three parts, equivalent to the stages described by Box and Jenkins (1976) [10]. In the identification stage, it outlines the response series and identifies candidate ARIMA models for it. And it analyzes time-series data that are to be used in later statements, possibly distinguishing them, and calculates autocorrelations, inverse autocorrelations, partial autocorrelations, and cross-correlations. Stationarity tests can be performed to determine if differencing is necessary [10].

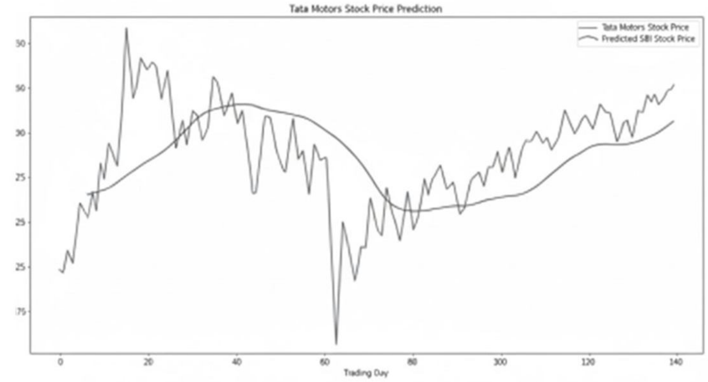


Figure 5. LSTM Regression Stock Price Prediction graph

to see if parts of the equation can be omitted for a more effective simpler model.

Journal of Development Economics and Management Research Studies (JDMS), A Peer Reviewed Open Access International Journal, ISSN 2582 5119 (Online), 09 (11), 55-66, January-March, 2022. 58 If the diagnostic tests demonstrates problems with the model, you try another model and then repeat the estimation and diagnostic checking stage [10]. In the forecasting stage, future values of time series are forecasted and to generate confidence intervals for these forecasts from the ARIMA model produced by the preceding estimating stage [10]

Long Short-Term Memory (LSTM) networks, a specialized form of recurrent neural networks (RNNs), have gained prominence for time series prediction, especially in financial markets, due to their capability to capture long -term dependencies and complex non -linear patterns.

Characterized by their gated memory cells, LSTMs can retain essential information over extended sequences, effectively addressing the vanishing gradient problem encountered in traditional RNNs. This feature makes them particularly suited for modelling stock price data, where past events can have delayed effects on future prices. Where LSTMs excel compared to many traditional approaches for stock price forecasting, however, is the capacity to more accurately develop non-linear relationships and discover complex lagged effects through time. However, LSTM models are not without their own limitations including extensive quality data required for training, intensive computational requirements, and sensitivity to specific hyperparameters (layer depth, learning rate, batch size). Furthermore, blackbox modeling lowers interpretability which implies that practitioners do not necessarily have the best understanding of what's causing the outcomes. [4][6][9].

Subsequent enhancements involve Bi-directional LSTMs and stacked LSTM networks which possess even higher predictive powers through reliance on forward and backward predictive relationships of financial sequences [6][9]. Frameworks that create a symbiosis between LSTMs and CNN layers (CNNLSTM) or attention mechanisms also work to improve feature learning and time step attention [30]. Yet interpretability problems and expensive computations limit LSTMs from functioning on a large scale in financial ecosystems and more recently, scholars have sought to alleviate this through XAI efforts, explainable AI with goals of substantiating end user predictive interpretation for LSTAbased projections [25][30]. The fact that LSTM is a common mechanism of financial study implies both its success as an innovative predictive mechanism but also, its dependence upon other models and outside sentiment/event analysis for the best practical, real world prediction modeling solution.

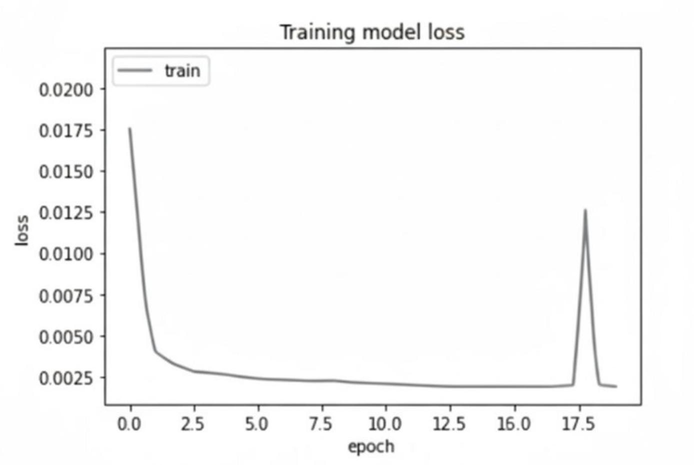


Figure 2. Loss vs. Epoch in ANN

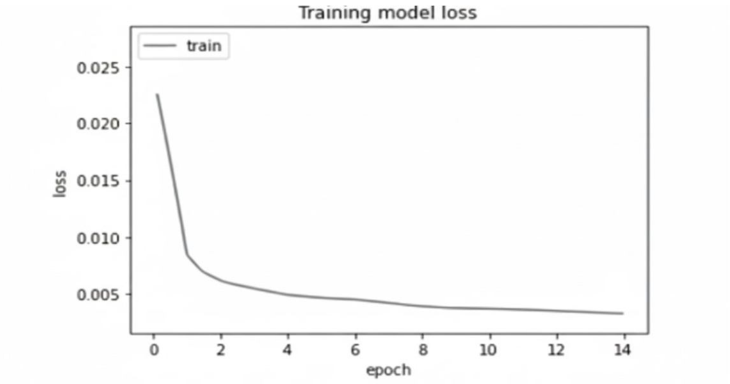


Figure 3. Loss vs Epoch in LSTM

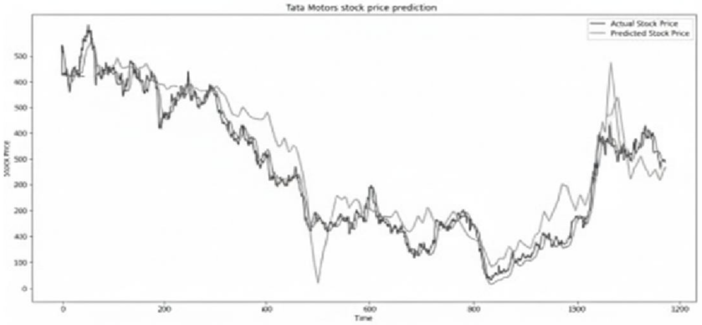


Figure 4. ANN model Tata Motors Stock Price vs Time

Error analysis of the model using residuals plots is illustrated in Figure 6. The graph on the left shows the data versus the prediction error and the

regressions that residuals scatter around the line at zero. The residue shows that the errors are quite small and distributed randomly without any coherent patterns.

This randomness indicates that the model does not suffer from systematic bias in its predictions.

The right side graph shows a histogram of the residuals which approximates a normal distribution.

This pattern further validates the hypothesis that residuals are random noise which increases the trust in the prediction model.

The combination of these plots shows that the model's predictions are not biased and are within the error margin of acceptance

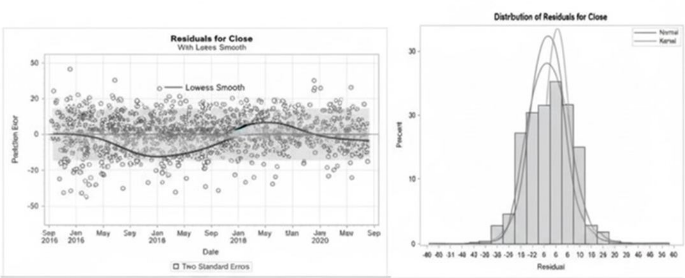
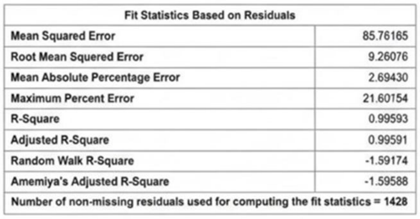


Figure 6. Data vs Prediction Error and Residual vs Percentage



Fit Statistics Based on Residuals	
Mean Squared Error	85.76165
Root Mean Squered Error	9.26076
Mean Absolute Percentage Error	2.69430
Maximum Percent Error	21.60154
R-Square	0.99593
Adjusted R-Square	0.99591
Random Walk R-Square	-1.59174
Amemiya's Adjusted R-Square	-1.59588
Number of non-missing residuals used for computing the fit statistics = 1428	

Figure 7. Statistics of UCM

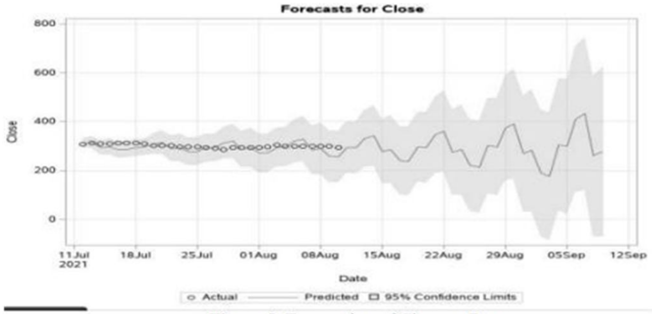


Figure 8. Forecasting of Close vs Date

The detail of UCM statistical evaluation is outline in Figure 7. It provides MSE, RMSE, MAPE, and R-Squared (R²) values. The small error values of MSE, RMSE, and MAPE suggest that UCM model predictions align closely with the existing stock prices. The R² and Adjusted R² values, as well as the other R² values being very high, closely 1, confirm that the model accounts for nearly all variation within the dataset and, as such, further substantiates According to the above graph, in the first, the loss of the model is so much its strength. These values help in assessing the predictive mastery of the that the prediction error is more, but after some epoch, the predicted values model in contrast with other forecasting methodologies such as ARIMA, is similar to the actual stock values. For this model, Adam optimization is LSTM, and complex hybrids, thus underscoring the model's utility. used. Adam optimization is a stochastic gradient descent optimization method that is based on the adaptive estimation of first-order and second-Figure 8 shows how stock closing prices have been projected over time. It order shows both actual and predicted stock closing prices, including a confidence interval of 95 percent. The close values of actual and predicted values

3.3.3. Introduction to LSTM Networks for Time Series Prediction suggest how well the model captures the trends. The confidence band shows the uncertainty range of forecasts and it is assured that the upper and lower values of the interval are expected to contain the actual values with a high probability. Such a model is useful for prediction - not just as a point estimate - because it takes into account the variation of actual values and the uncertainties of actual values, which are two critical components for estimations in financially unpredictable situations. Thus, the figure substantiates the relative accuracy of the model based upon stock price assessment expectations. Ultimately, the predictions made by the model are empirically beneficial and are easier to comprehend Finally, a few disadvantages of LSTMs exist. For example, LSTMs face challenges from optimal sequence and sequence length; too many/malformed sequences result in overfitting and predictive failure. Therefore, regularization techniques - dropout, cross-validation, and early stopping - are necessary to prevent reduced generalizability when implemented in practical scenarios [19]. Furthermore, training LSTMs is also a time-intensive process that requires significant time with a powerful computing setup (advanced GPUs) with large datasets to provide the most accurate predictions. However, despite such pitfalls, the fact that LSTMs can operate successfully in a vague and nonlinear environment makes them a favored deep learning method for finance-related endeavors. For example, much literature suggests that LSTM outperforms ARIMA and other traditional time series approaches for low-frequency predictions for less volatile, long-term trends and more high-frequency shocks [6][9][25].

D. Comparative Studies Between ARIMA and LSTM in Stock Prediction

Here we have compared the ARIMA and LSTM model on parameters such as on error metrices and Data window size.

1) Performance Comparison on Error Metrics

Comparative analyses between ARIMA and LSTM models consistently According to these standard prediction performance metrics, LSTM is the best in terms of prediction performance.. For instance, a study conducted on SP500 stocks reported that LSTM networks achieved up to a 92.

2) Data Window Size and Impact on Model Efficiency

The implementation of ARIMA and LSTM models is strongly influenced by the size and characteristics of the historical data window used for training. ARIMA models generally exhibit better results with long, stationary time series, as they leverage their statistical assumptions and effectively utilize stable autocorrelations present in extended datasets. In contrast, LSTM models often struggle with smaller datasets due to their dependence on large amounts of data for accurately learning temporal dependencies and complex dynamics. When the data is weakly stationary or demonstrates limited volatility, ARIMA’s linear framework and statistical interpretability provide a clear advantage. However, as the dataset expands in length and complexity, LSTM architectures become more effective, owing to their ability to capture nonlinear behaviors and adapt to rapidly changing market conditions
[7][8][11].

Moreover, the choice of data window size directly impacts computational efficiency and forecasting horizons. Shorter windows tend to reduce training time and prevent overfitting in ARIMA but may limit the depth of patterns captured. LSTMs, on the other hand, benefit from longer windows, where extended historical sequences provide richer temporal context for learning. However, excessively large windows can introduce noise and redundant information, which may degrade performance if not managed through feature selection or regularization. Thus, selecting an optimal data window size becomes a trade-off between accuracy, generalization, and efficiency, with the ideal configuration varying across datasets and market scenarios [6][25].

3) Hybrid and Ensemble Approaches Combining ARIMA and LSTM

In relation to the two studies above, several subsequent studies have investigated hybrid models based on the strengths of ARIMA and LSTM. One hybrid configuration is to assess time series with ARIMA for linear components and assess the residuals, and only then, intervene with LSTM networks for non-linear relationships in the unknown residuals. Thus, this hybrid modeling approach can account for stationary components and complicated temporal relationships better than one single model as it gets the complete picture.

As a result, such hybrid designs boast improved predictive accuracy, championing stronger more generalized results with reduced error metrics compared to developing each model independently with ARIMA or LSTM. For example, the following study notes that in addition to fine tuning and early stopping based on parameter modifications, generalization and accuracy were substantially increased, championing the benefits of ensemble hybrid models for stock price forecasting performance. [18][22][34].

E. Applications of ARIMA in Stock Price Forecasting

Why ARIMA for stock price prediction because the model is capable of capturing past prices and Autoregresssion(AR) defines how it influences the future prices. ARIMA can still be applied to capture short term patterns and trends in price movements. ARIMA is efficient for short term forecasting.

1) Traditional Use of ARIMA in Various Markets

ARIMA is frequently used for stock market and commodity market; forex/currency market; bond and interest rate market; energy market, among others.ARIMA models remain highly relevant in practical financial forecasting, particularly for short-term stock price predictions in markets such as the NYSE, NSE, and SP 500. The modeling process generally begins with statistical validation of stationarity using tests like the

Augmented Dickey-Fuller (ADF), followed by model identification and selection guided by information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Subsequently, the forecast performance is assessed using error statistics of ensample and out-of-sample data to validate model stability. Such approaches have been used successfully to forecast index level data of as well as certain stock prices, for investors/analysts in due time to better make data driven financial decisions and grasp the immediate phenomenon.[12][13][15].

2) ARIMA Model Enhancements for Improved Accuracy

Recent advancements aim to enhance the predictive capabilities of ARIMA by incorporating sophisticated techniques such as regularized Gaussian basis function expansions, leading to generalized linear modeling frameworks. These improvements enable capturing functional data features without restrictive smoothing assumptions, addressing some of ARIMA's limitations in handling high-dimensional time series data. Seasonal extensions like SARIMA models integrate seasonality effects critical in certain stock price series. Additionally, combining ARIMA with other statistical frameworks like GARCH (Generalized Autoregressive Conditional Heteroskedasticity) helps model volatility clusters, thereby improving portfolio risk assessment and price prediction accuracy [26][27][15].

3) Limitations of ARIMA in Stock Price Prediction

ARIMA is robust and therefore popular, but the simplistic nature of its linear derivations creates fundamental shortcomings. For instance, it fails to acknowledge nonlinear developments that punctuate a number of financial markets - erratic surges in price, regime shifts based on globalization's economic/political cause and effect, etc. Thus, ARIMA is more prone to forecasting error during turbulent times and globalized systemic failure. Furthermore, it uses historical observations with the power of stationarity assumptions [14][21][25].

4) FinBERT model

The textual components of news can be used to predict short term stock movement. Sentiment analysis of news headlines can be crucial in assessing stock performance in the short term. Bidirectional Encoder Representations from Transformers (BERT) is one of the leading Artificial Intelligence models in Natural Language Processing (NLP) which is developed by Google. BERT is trained in both directions which is why it is able to comprehend a certain context better than the regular models from the left or the right side. BERT contains is built with transformers which learns contextual associations between words in a body of text. BERT has been subjected to pre-training based on two tasks, Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). An original BERT paper contains comprehensive discussion on the model architecture, pre-training and fine-tuning.

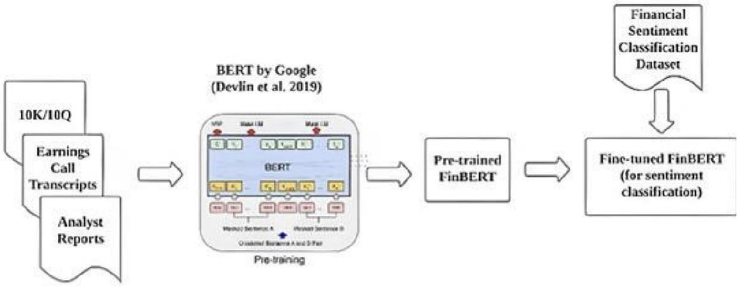

10K/10Q
Earnings Call Transcripts
Analyst Reports
BERT by Google (Devlin et al. 2019)
BERT
Pre-training
Pre-trained FinBERT
Financial Sentiment Classification Dataset
Fine-tuned FinBERT (for sentiment classification)

Figure 9. FinBERT pretraining and fine-tuning (for sentiment classification)

F. Applications of LSTM in Stock Price Forecasting

LSTM is used for prediction of future stock prices and mainly for trend prediction will market go up or down and volatility forecasting and multi asset forecasting these are the feils where LSTM is used and LSTM helps in capturing nonlinear patterns.

1) LSTM Architectures Applied to Stock Prediction

In addition, hybrid architectures that integrate convolutional neural networks with LSTM (CNN-LSTM) have been introduced to simultaneously extract spatial features from financial indicators and capture long-term temporal dependencies, yielding superior performance in multivariate forecasting scenarios [30][31]. Attention-based LSTM extensions have further advanced this domain by enabling the model to assign greater weight to critical time steps, improving interpretability and accuracy in volatile markets [6][33]. Recent work has also experimented with ensemble models that combine LSTM with Gated Recurrent Units (GRUs) or Transformer layers, providing complementary strengths in capturing short- and long-term dependencies. Such integrations have been shown to outperform single deep learning models by balancing efficiency, accuracy, and generalization in dynamic market environments [30][33]. Furthermore, transfer learning techniques using pre-trained LSTM frameworks are gaining attention, as they allow leveraging knowledge from one financial domain to another, thereby reducing training time and improving adaptability in cross-market forecasting tasks [6]. Collectively, these advancements confirm LSTM as a foundational architecture in financial deep learning, while highlighting ongoing innovations that continue to push the boundaries of stock market prediction.

2) Datasets and Features Utilized in LSTM Models

LSTM models typically employ historical price data as primary input features, including open, high, low, close prices, and volume. Additionally, advanced approaches have integrated technical indicators derived from price and volume data to provide richer contextual information. Sentiment analysis derived from financial news headlines and social media has been combined with price series to incorporate external market sentiment fac- tors, improving predictive performance. Studies have also explored training on multiple stocks or indices simultaneously, such as the SP 500 and Indonesia’s LQ45 financial sec- tor indices, thereby showing the flexibility and generalizability of LSTM models across diverse markets and time horizons [28][29][30].

3) Performance Evaluation of LSTM Models

Analyses consistently show that LSTM models outperform traditional statistical models in capturing volatile and rapidly changing stock price patterns, given their ability to dynamically model non-linear relationships and long-term dependencies. The use of techniques such as early stopping and dropout regularization mitigates the risk of overfit- ting and improves model generalizability. However, LSTM training demands considerable computational resources, and the model’s accuracy heavily depends on carefully tuned hyperparameters. Optimizing these factors effectively improves prediction stability and accuracy, with empirical evidence confirming enhanced performance on error metrics across multiple datasets [9][19][20].

IV. ERROR METRICS AND EVALUATION METHODS

The both of the models are evaluated on evaluation metrices and evaluation methods. The various metrices used for evaluation.

A. Common Metrics Used in Stock Prediction Evaluations

The evaluation of forecasting models in stock price prediction predominantly relies on error metrics such as RMSE, MAE, and MAPE. RMSE emphasizes larger errors by squaring residuals, making it suitable for assessing models where large deviations are critical. MAE provides a linear measure of average absolute errors, serving as a more interpretable but less sensitive metric. MAPE, expressed as a percentage, allows for comparison across stocks with different price scales. Validation techniques like walk- forward validation, where models are retrained progressively with new data, and cross- validation are widely adopted to ensure robustness and assess generalization. The choice of error metric often aligns with specific forecasting objectives and the relevant prediction horizon [31][32][33].

B. Statistical Validation in ARIMA Modeling

Model assessment in ARIMA systems supports statistical testing for ado- quaky. This is important because selection is based on the AIC and BIC of commonly selected order estimations which ensure the best parameterized system without added confusion and error. The Ljung-Box test for residual diagnostics determines if residuals are white noise (no autocorrelation exists) which means a correctly configured model has been developed. These empirically support an ARIMA since tests for stationarity - Augmented Dickey Fuller (ADF) and Phillips-Perron (PP) - support that before determining that enough differencing rendered the series stable.

These empirical findings are critical to subsequent confirmation of the ARIMA models structural assumptions for forecasting. [12][13][15].

C. Deep Learning Model Validation for LSTM

LSTM models incorporate validation techniques such as early stopping to halt training once the model's performance on a validation set ceases to improve, effectively preventing overfitting. Regularization methods like dropout randomly deactivate neurons during training to foster model robustness. Hyperparameter tuning, achieved through grid search or more advanced methods, critically influences model accuracy. The selection of train-test splits and the proportion of data allocated to each partition considerably affect model performance and generalizability.

Strengths and Weaknesses of ARIMA vs LSTM Mod**els**

Here the comparison between the both models is shown where the strengths and weakness are discussed.

D. ARIMA Model Strengths

ARIMA's notable strengths include robustness in small datasets, particularly where the time series is stationary or exhibits linear behavior. Its clear parametric structure delivers transparency and interpretability, allowing analysts to understand the contribution of individual parts of the model such as lag orders and moving average terms. The extensive theoretical foundation and well-developed software implementations make ARIMA a preferred choice in many traditional forecasting contexts, including finance. For practitioners prioritizing explainability and short-term prediction scenarios with limited data, ARIMA remains a valuable tool [12][13][15].

E. LSTM Model Strengths

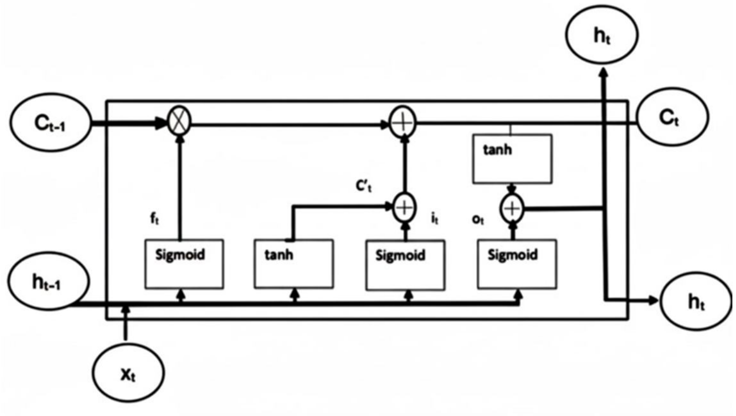


Figure 10. A Long Short-Term Memory network.

Conversely, LSTM models excel at modeling complex temporal dependencies and nonlinear dynamics often exhibited in financial time series. Their memory gating mechanisms enable retention and selective forgetting of long-range temporal information, critical for markets where past influences extend over varying time scales. The adaptive learn- ing capability allows LSTM models to adjust to changing market regimes and volatility structures. These qualities facilitate superior performance in volatile and non-linear en-virements, enhancing forecasting accuracy across diverse market conditions [6][7][9]

F. Limitations of Both Models

While both models were great with prediction feasibility of the non-linearity and model potential forecasting stock prices that practitioners should tread lightly after awareness of limitations, these two models have the following limitations that forecasting practitioner should be aware of relative to these two:

For ARIMA, the limitations are: 1) it's linear, therefore linearization of nonlinear behaviors must occur; 2) as a non-linear model, LSTM will fail in crises where non-stationarity and structural breaks exist and impact the situation more powerful than LSTM can learn from what's within the model. However, relative to LSTM, additional factors emerge that must be considered: 1) ARIMA LSTMs require a lot more resources - computers, data preprocessing; 2) LSTMs become overfit in the testing and validation if hyperparameter tuning is not handled properly; 3) black box features increase uncertainty in any interpretability which bad for the financial trading world; 4) LSTMs require stationarity which is not reality in finance; this means too much differencing or transformation periods occur which breaks up realism of potential relationships; 5) this is compounded by ARIMA who does not handle high frequency trading data well either where noise and wild volatility mess everything up; and 6) ARIMA is short term predictive which means it's not able to gain empirical insights substantiated for practical application for investments because it doesn't account for time horizons. Relative to LSTM, however, relative to deep learning architectures which are great because they allow accommodation of non-linearities LSTMs are problematic in fields where there's not enough history or financial capital (emerging markets). Thus. LSTMs become problematic. In addition, training is computational intensive; LSTM needs GPU support which means longer training times which smaller firms with access to this training get access. Thus, smaller companies would not be able to get the applied benefits from implemented successful LSTM operations without exception. Finally, performance is very dependent on hyperparameters - window size, learning rate, dropout rate; without proper fine tuning, variable results come with unstable or biased predictions. [9][20].

V. ADVANCES IN HYBRID AND ENSEMBLE MODELS INCORPORATING ARIMA AND LSTM

In recent years,researtchers have developed hybrid models that combine the strengths of traditional statistical methods and modern Machine learning/deep learning to improve stock price forecasting.

A. Hybrid Model Architectures and Their Rationale

Hybrid models that combine ARIMA and LSTM aim to leverage the linear modeling strengths of ARIMA with the nonlinear learning capabilities of LSTM. Typically, ARIMA first models' linear components and residual errors, which are then passed to LSTM networks that learn the remaining nonlinear patterns. Where hybrid models take the mean of the forecasts of the component models employed, ensemble approaches take the average of the forecasts from each, as composite forecast reliability and accuracy even more so because it decreases the forecast error of the component models. Composite approaches were studied more over time as viable combinations that suitably balance the complex, often chaotic characteristics of stock market movements because over time, the limitations of single approaches have been acknowledged. [18][22][34].

B. Performance Gains and Practical Implications

Empirical findings indicate that those models that combine ARIMA or LSTM with another model achieve more accurate forecasts and lower error when comparing each model independently. For example, the decreased RMSE and MAPE statistics demonstrate the operational effectiveness of hybridized forecasting from both ARIMA and LSTM, especially for erratic markets featuring linear and nonlinear events. The more flexible the combinations, the better they can function across different market regimes and securities. Therefore, performance improvements are advantageous to real-world functioning in such areas as algorithmic trading, portfolio risk management and strategic forecasting where forecasts must be accurate and given in a timely manner. [34][35][36].

C. Challenges and Limitations of Hybrid Approaches

Despite their promising results, hybrid models also bring forth several challenges that limit their widespread adoption in financial forecasting. The integration of multiple models substantially increases computational complexity, often requiring greater processing power, memory resources, and longer training times compared to standalone ARIMA or LSTM models.

D. Feature Selection and Data Preprocessing

There are two main attribute characteristics that influence model efficacy relate to data transformation and feature selection. For example, one could combine technical indicators, volume and even sentiment values from news/social media to create a three-dimensional input which, potentially, from a common sense perspective, gives a model more supportive contextual features. In addition, normalization and smoothing of values would create stable distributions over time making a model more likely to achieve convergence when trained off such values. Ultimately, the inclusion of macroeconomic indicators and external forces compensate for the limitations of purely price driven historical regression predictions.

The inclusion of sentiment scores derived from news articles, analyst opinions, and social media platforms such as Twitter and Reddit has gained significant traction in stock prediction research. Natural Language Processing (NLP) techniques, including sentiment analysis and event extraction, provide additional signals that capture investor psychology and market mood, which traditional quantitative features may overlook [6][33]. By integrating these qualitative dimensions with numerical data, predictive models become more robust against sudden shifts caused by market rumors, announcements, or geopolitical events.

Furthermore, many preprocessing data measures ensure model accuracy. For instance, normalization (Min-Max, Score standardization) prevents any one feature from being more impactful than others - especially those features which are predisposed to higher volumes. In addition, smoothing - moving average, exponential smoothing - prevents a model from responding too much to a volatile series and instead prioritizes historical price action over random noise.

Moreover, stock price movements are influenced by macroeconomic (and exchange traded derivatives) factors from interest rates to inflation to GDP growth to currency exchange. Thus, exogenous variables that provide a context for stock price prediction can align models better with business cycles. Exogenous variables are more universally applicable across different markets while risk exogenous variables (VIX, oil and gold prices) represent inter-market relationships. [25][31].

E. Impact of Dataset Characteristics

Model performance is significantly influenced by dataset properties such as length, frequency, stationarity, and the presence of structural breaks. ARIMA models benefit from longer, more stable datasets where stationarity can be enforced. However, market volatility and abrupt regime changes undermine prediction accuracy for both ARIMA and LSTM. LSTM, while more flexible, is affected by missing data and noise, which require preprocessing methods to maintain model robustness. Managing these dataset char- act eristics is essential to harness each model’s strengths effectively and ensure reliable forecasting outcomes [14][19][25].

F. 5Model Training Strategies and Parameter Optimization

Effective training strategies and parameter tuning are vital for attaining model accuracy and generalization. In LSTM, hyperparameter optimization via grid or random search and early stopping techniques prevent overfitting and directional convergence. As for ARIMA, parameters are usually set via AIC measurement, in- sample identification and tweaking. Cross-validation assesses whether it's overfitted and tests stability so that researchers can establish a stable model amid differing data scenarios. [9][13][20].

G. Integration with Other Deep Learning Models

Industry developments involve LSTM combinations with other deep learning Architectures for enhanced prediction. For example, hybrid CNNLSTM networks take advantage of convolutional layers to perform feature selection first and then apply sequential pattern recognition to increase accuracy and speed of processing. There are also bidirectional LSTMs that incorporate information from the sequence prior and post to the present point of interest. Attention and transformer models more prevalent in NLP are also considered in financial time series for dependency ascertainment and multistep forecasting horizons [28][29][30].

H. Incorporation of Alternative Data Sources

Other areas for potential future research relative to the findings of this study stem from more contemporary empirical research in the field to date beyond price and volume data, much of contemporary empirical research in the field to date concerns supplementary alternative data like news/social media sentiment data, order book data, high frequency trading data. These are more modern offsets of sentiment data and microstructure data that have a need for projection and which need more time sensitive predictions. In addition, in an attempt to bridge gaps of datasets for modeling when real world data cannot be consistently found or found at all, utilizing Generative Adversarial Networks (GANs) is a major opportunity for future research. [2][5][37].

Sentiment data relative to social media - Twitter, Reddit or financial news - proved even more effective relative to sentiment and herd mentality. For instance, it's been noted that social media sentiment data insinuates intraday or price levels before confirmed volume and price findings [6][33]. In other words, tonality, positive/negative, gives modelers another factor to predict with beyond historical numeric patterns. Therefore, social media sentiment increases predictability.

The same applies to order book data and high frequency trading data. These two types of data less frequently manipulated represent market microstructure via bid-ask spread and liquidity depth in addition to the speed with which trades are executed representing investor/institutional interest. Thus, the more these characteristics are used in a model, the more predictable gaps or liquidity shocks are before they're recognized by daily closing price evaluations.[19][31].

Furthermore, increasingly more information enters into the hands of the party, like google searches, corporate filings and ESG ratings for variables that position market sentiment and longer term risks into consideration which, in fact, are already operating. These variables are less contestable through predictive modeling because they function on a shorter time frame for more responsive industries based upon consumer sentiment and/or sociopolitical sentiment. [25].

VI. ADDRESSING MODEL INTERPRETABILITY AND ROBUSTNESS

Interpretability refers to ability of diverse models and data pipelines to work together seamlessly, ensuring consistent and key full outputs.

Addressing the black-box nature of deep learning models remains a priority, with efforts directed at explaining LSTM predictions in financial contexts. Techniques such as SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explain- nations) are being adapted to make model outputs more transparent for practitioners. Furthermore, improving robustness against market shocks, regime changes, and Incorp- rating uncertainty measures are active research areas aimed at creating reliable systems suitable for deployment in volatile real-world environments. Combining statistical rigor from models like ARIMA with the flexibility of AI methods represents a promising strategy [25][30][37].

VII. SYNTHESIS OF KEY FINDINGS

The third part discusses the advantages and disadvantages of the ARIMA and deep learning models.

To conclude, LSTM is a more appropriate forecasting model because it provides a more accurate forecast with internalized time dependencies of the data used, as per the historical nature of the test, to lessen forecasts corresponding to nonlinear, nonstationary characteristics of the forecast. This does not mean that ARIMA is refuted as an accurate model for certain forecasting conditions (linear, stationary, short-term) but instead fails in some situations. In addition, both LSTM and ARIMA hybrid models show that the two models learn from each other - the LSTM accuracy increases while the ARIMA feasible solutions coincide with the nature of all financial markets.[6][13][34]

VIII. IDENTIFIED RESEARCH GAPS AND CHALLENGES

Such gaps and shortcomings necessitate increasingly sophisticated, explainable and comprehensive prediction models boasting a heterogeneous approach, flexibility to market fluctuations and degrees of implementation across settings.

While significant developments have been made, there are a number of future research prospects due to comparability metrics and lack of a homogeneous dataset that render cross-the-board analysis improbable. Many authors rely on varying timeframes, stock indices and geographically specific markets that render fair distribution for the interested models improbable while rendering reproducibility complicated [19][25]. In addition, models are only somewhat explainable in comparison to deep learning, and as a domain where explanation is necessary for persuasion and compliance, few incorporate SHAP values and LIME and attention mechanisms in stock prediction fields [36].

The fourth limitation is relative homogeneity of the market. Much is from generalizable, larger markets like the SP 500, NASDAQ or NIFTY; few, for instance, explore emerging markets, commodities or sector specific stocks which both make findings less generalizable but fails to acknowledge different financial ecosystems that respond heterogeneously. The same idea holds for exogenous variables - only a small few investigate macroeconomic shocks, geopolitical events or policy announcements as a predicable variable and many real world factors that influence data remain uninvestigated.[19][25]

Furthermore, few explore high frequency trading dynamics or intra-day prediction where access is more heterogeneous and difficult in addition to much more noise and information that is not as established and permanent. Yet these fields need established models and dynamic learning for predictive performance despite access issues. This holds true for out of sample predictions greater than one year in advance - however fewer documents go this far - although out of sample predictions in month increments have even less coverage than daily in addition to hybrid models needing more coverage with consistent reliability across studies assessing to see if hybridization helps based on modeling or data available.

IX. RECOMMENDATIONS FOR PRACTITIONERS AND RESEARCHERS

For practitioners, adopting hybrid ARIMA-LSTM frameworks with rigorous validation protocols is advised to balance interpretability and predictive power.

Researchers should continue exploring novel architectures, including transformers and attention mechanisms, while integrating alternative data such as sentiment and high-frequency indicators. More- over, enhancing model explainability and robustness is crucial to bridge the gap between academic advances and real-world deployment, ensuring reliable, transparent, and trust- worthy stock prediction systems [40][41][42].

X. METHODOLOGY

The section describes the selection, preprocessing and transformation of datasets to ensure reliability.

As long as scientists and investors alike seek a return above the one available via stock market investment, stock price prediction is set to become an ever more researched field [5,6]. The primary goal of investors is an accumulation of wealth from stock market investment. Therefore, from these findings and the run/results of ARIMA, ANN and UCM, this study implies that in terms of MSE that UCM is best because it encompasses seasonality, and in light of this, LSTM is next up due to the ANN. However, in terms of ARIMA, it falls short and needs higher accuracy through a better white noise test. Yet the percentage error for LSTM is much less than that of ARIMA and UCM. The only negative thing about stock price prediction is that it won't change unless the firm changes it, but many economic and socio-political factors, investments by the firm, current events, and new IPOs by competing firms will influence stock price. Yet these three models predict without this information. More accurate predictions can be made when these things come into play, as well as the yearly company budget for sentimental analysis prediction. For example, a new model can be developed through training with ANN and predicting stock price with ARIMA while applying the sentiment analyses.

Figure 11 indicates the actual Apple's stock prices and the corresponding predicted values using LSTM and ARIMA models. The predicted values using LSTM are in good agreement with the actual trend whereas ARIMA seems to have larger variances. This emphasized again the greater capability of LSTM to capture the non-linear movements of the stocks.

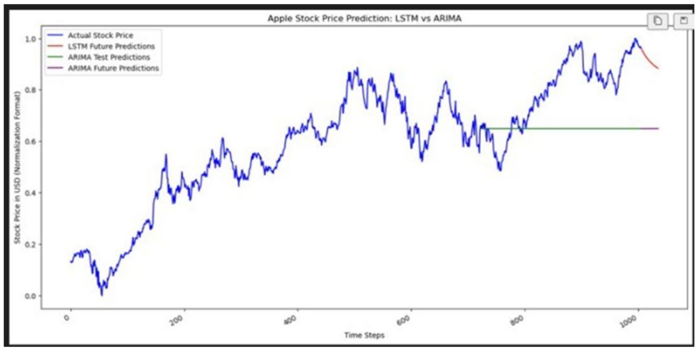


Figure 11. Apple Stock Price Prediction: LSTM vs ARIMA

Figure 12 predicts and forecasts the stock prices of Apple for the following days using LSTM. The stock prices that are predicted (in red) match the actual stock prices in the sense that they follow the pattern of stock prices going up and down. This shows that predicted values of LSTM are consistent which are beyond the actual observed values in data and fall within the trend.

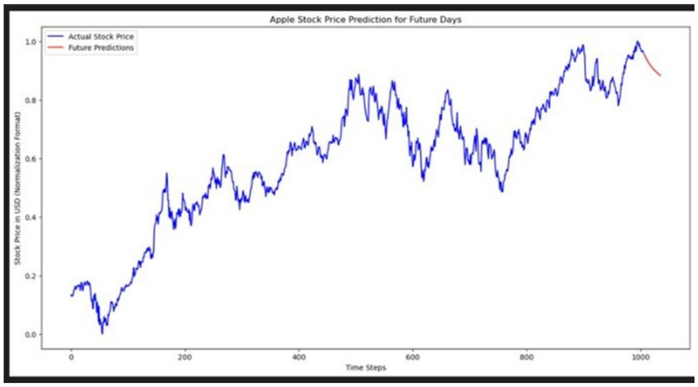


Figure 12. Apple Stock Price Prediction for Future Days

XI. CONCLUSION

I have used multiple python libraries, mainly pandas, to analyze and visualize data pertaining to stocks, especially technology stocks, and their respective data from the stock market. Also, this paper attempts to analyze the stock risk based on the past performance and uses stock price prediction using LSTM model and ARIMA model. The historical dataset available on the company's website is lacking in several aspects as it only covers a few fundamental pillars such as high and low stock prices, closed and opened stock prices as well as trading volumes. In order to augment accuracy, additional variables are generated from the features.

LSTM model experiment on Apple stock price, 95. Rather than calculating a simple moving 95 average, which rests on the premise of calculating the average of the last N values, the model incorporates the use of several randomly selected short subsequences from the training dataset. The method df[col]. rolling(N), which corresponds to the command used to create a rolling window, applies the same principle and helps in the generation of a window of size N for each timestamp t such that the outputs are the rows t, t-1,..., t-(N-Ver1), and t the set N is the number of the rows to be shifted. This method of filling aims to keep the order of the inputs. The inputs for the t timestamp are obtained after the values have been shifted prediction and the last value is set to NaN as is required the. The last step is to compute the predicted values with…. From Figure 11 and Figure 12, it can....

REFERENCES

[1] Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383–417.

[2] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. Wiley

[3] Chatfield, C. (2003). The Analysis of Time Series: An Introduction. CRC Press.

[4] Ariyo, A., Adewumi, A. and Ayo, C., (2014). Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. Gers, F. A., Schmidhuber, J., Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. Neural Computation, 12(10), 2451–2471.

[5] Agrawal, N., (2019). Stock Market Prediction Approach: An Analysis. International Journal of Engineering Research & Technology (IJERT), 06(03), pp.847-849.Nelson, D. M., Pereira, A. C., de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. International Joint Conference on Neural Networks (IJCNN).

[6] Haider Khan, Z., Sharmin Alin, T. and Hussain, A., (2011). Price Prediction of Share Market Using Artificial Neural Network 'ANN'. International Journal of Computer Applications, 22(2), pp.42-47.

[7] Fischer, T., Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654–669.

[8] Siami-Namini, S., Tavakoli, N., Namin, A. S. (2019). A comparison of ARIMA and LSTM in forecasting time series. IEEE International Conference on Machine Learning and Applications (ICMLA).

[9] 0. Support.sas.com. (2021). PROC ARIMA. [online] Available at: [Accessed 20 August 2021].

[10] Shynkevich, Y., McGinnity, T. M., Coleman, S. A., Belatreche, A. (2017). Forecasting price movements using technical indicators: Investigating the impact of varying input window length. Neurocomputing, 264, 71–88.

[11] Fama, E.F. Efficient capital markets: A review of theory and empirical work. J. Financ. 1970, 25, 383–417. [CrossRef] 2. Malkiel, B.G. Efficient market hypothesis. In Finance; Springer: Berlin, Germany, 1989; pp. 127– 134.

[12] Deorukhkar, O., Lokhande, S., Nayak, V. and Chougule, A., (2019). Stock Price Prediction using combination of LSTM Neural Networks, ARIMA and Sentiment Analysis. International Research Journal of Engineering and Technology (IRJET), 06(03), pp.3497-35003..

[13] Zhang, G., Eddy Patuwo, B., Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35–62.

[14] Peng, H., Zhou, Y. (2022). Hybrid ARIMA–LSTM models for financial time series forecasting: Enhancing robustness and interpretability. Applied Soft Computing, 114, 108108.

[15] Liu, H., Chen, L., Xu, Y. (2022). Attention-based deep learning models for stock prediction: A survey. Neurocomputing, 489, 336–356.

[16] Xing, F., Cambria, E., Welsch, R. E. (2020). Natural language based financial forecasting: A survey. Artificial Intelligence Review, 54, 3763– 3812.

[17] Wu, H., Zhang, J. (2021). Hybrid ARIMA–LSTM model for stock price prediction. IEEE Access, 9, 33011–33020.

[18] Kara, Y., Boyacioglu, M. A., Baykan, O. K. (2011).¨ Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert Systems with Applications, 38(5), 5311–5319.

[19] Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press. [20] Zhao, K., Zhang, C., Li, H. (2019). Event detection from social media data streams. IEEE Transactions on Knowledge and Data Engineering, 31(7), 1234–1248.

[21] Atkinson, P., Campos, L. (2020). Event detection and classification in heterogeneous data. Springer Lecture Notes in Computer Science, 12013, 112– 128.

[22] Chen, Y., Wang, S., Zhai, X. (2021). Deep learning for event extraction: A survey. ACM Transactions on Knowledge Discovery from Data, 15(4), 1– 33.

[23] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan Claypool Publishers.

[24] Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), 102–107.

[25] Zhang, L., Wang, S., Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

[26] Li, J., Sun, A., Ma, J. (2021). Event-driven sentiment analysis: Methods and applications. IEEE Access, 9, 75712–75729.

[27] Xu, Y., Cohen, S., Zhao, T. (2020). Temporal sentiment-event analysis for financial markets. Information Processing Management, 57(3), 102256.

[28] Schwert, G.W. Why does stock market volatility change over time? J.
Financ. 1989, 44, 1115–1153

[29] Chan, J.Y.L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.W.; Chen, Y.L. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics 2022, 10, 1283

[30] Chan, J.Y.L.; Leow, S.M.H.; Bea, K.T.; Cheng, W.K.; Phoong, S.W.; Hong, Z.W.; Chen, Y.L. A Correlation-Embedded Attention Module to Mitigate Multicollinearity: An Algorithmic Trading Application. Mathematics 2022, 10, 1231

[31] Li, Q.; Wang, T.; Li, P.; Liu, L.; Gong, Q.; Chen, Y. The effect of news and public mood on stock movements. Inf. Sci. 2014, 278, 826–840. [CrossRef] 16. Jiang, W. Applications of deep learning in stock market prediction:
Recent progress. Expert Syst. Appl. 2021, 184, 115537.

[32] Ozbayoglu, A.M.; Gudelek, M.U.; Sezer, O.B. Deep learning for financial applications: A survey. Appl. Soft Comput. 2020, 93, 106384. [CrossRef]

[33] 18. Chopra, R.; Sharma, G.D. Application of Artificial Intelligence in Stock Market Forecasting: A Critique, Review, and Research Agenda. J. Risk Financ. Manag. 2021, 14, 526. [CrossRef] 19

[34] . Shah, D.; Isah, H.; Zulkernine, F. Stock market analysis: A review and taxonomy of prediction techniques. Int. J. Financ. Stud. 2019, 7, 26. [CrossRef] 20.

[35] Shahi, T.B.; Shrestha, A.; Neupane, A.; Guo, W. Stock price forecasting with deep learning: A comparative study. Mathematics 2020, 8, 1441.

[36] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," The Journal of Finance, vol. 25, no. 2, pp. 383–417, May 1970.

[37] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[38] C. Chatfield, The Analysis of Time Series: An Introduction, 6th ed. Boca Raton, FL, USA: CRC Press, 2003.

[39] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," Neural Computation, vol. 12, no. 10, pp. 2451–2471, 2000.

[41] V. Chang, P. Baudier, H. Zhang, Q. Xu, J. Zhang, M. Arami, and H. P. Le, "Financial data forecasting with deep learning: A systematic literature review (2010–2023)," Information Fusion, vol. 102, art. 101881, Jan. 2024.

[42] D. M. Nelson, A. C. Pereira, and R. A. de Oliveira, "Stock market's price movement prediction with LSTM neural networks," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Anchorage, AK, USA, 2017, pp. 1419–1426.

[43] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," Expert Syst. Appl., vol. 42, no. 1, pp. 259– 268, Jan. 2015.

[44] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," Eur. J. Oper. Res., vol. 270, no. 2, pp. 654–669, Oct. 2018.

[45] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Orlando, FL, USA, Dec. 2018, pp. 1394– 1401.

[46] S. Borovkova and I. Tsiamas, "An ensemble of LSTM neural networks for highfrequency stock market classification," J. Forecast., vol. 38, no. 6, pp. 600–619, Sep. 2019

[47] Y. Shynkevich, T. M. McGinnity, S. A. Coleman, and A. Belatreche, "Forecasting price movements using technical indicators: Investigating the impact of varying input window length," Neurocomputing, vol. 264, pp. 71–88, Nov. 2017.

[48] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of ARIMA and artificial neural networks models for stock price prediction," J. Appl. Math., vol. 2014, art. ID 614342, pp. 1–7, 2014.

[49] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," Int. J. Forecast., vol. 14, no. 1, pp. 35–62, 1998.

[50] H. Peng and Y. Zhou, "Hybrid ARIMA–LSTM models for financial time series forecasting: Enhancing robustness and interpretability," Appl. Soft Comput., vol. 114, art. 108108, Jan. 2022.

[51] H. Liu, L. Chen, and Y. Xu, "Attention-based deep learning models for stock prediction: A survey," Neurocomputing, vol. 489, pp. 336–356, May 2022.

[52] F. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: A survey," Artif. Intell. Rev., vol. 54, pp. 3763–3812, Mar. 2020.

[53] H. Wu and J. Zhang, "Hybrid ARIMA–LSTM model for stock price prediction," IEEE Access, vol. 9, pp. 33011–33020, 2021.

[54] Y. Kara, M. A. Boyacioglu, and O. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," Expert Syst. Appl., vol. 38, no. 5, pp. 5311–5319, May 2011.

[55] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[56] K. Zhao, C. Zhang, and H. Li, "Event detection from social media data streams," IEEE Trans. Knowl. Data Eng., vol. 31, no. 7, pp. 1234–1248, Jul. 2019.

[57] P. Atkinson and L. Campos, "Event detection and classification in heterogeneous data," in Lecture Notes in Computer Science, vol. 12013. Cham, Switzerland: Springer, 2020, pp. 112–128.

[58] Y. Chen, S. Wang, and X. Zhai, "Deep learning for event extraction: A survey," ACM Trans. Knowl. Discov. Data, vol. 15, no. 4, pp. 1–33, Jul. 2021.

[59] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool, 2012.

[60] E. Cambria, "Affective computing and sentiment analysis," IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar.–Apr. 2016.

[61] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, art. e1253, Jul.–Aug. 2018.

[62] J. Li, A. Sun, and J. Ma, "Event-driven sentiment analysis: Methods and applications," IEEE Access, vol. 9, pp. 75712–75729, 2021.

[63] Y. Xu, S. Cohen, and T. Zhao, "Temporal sentiment-event analysis for financial markets," Inf. Process. Manage., vol. 57, no. 3, art. 102256, May 2020.

[64] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," Expert Syst. Appl., vol. 42, no. 24, pp. 9603–9611, Dec. 2015.

[65] L. Nemes and A. Kiss, "Prediction of stock values changes using sentiment analysis of stock news headlines," J. Inf. Telecommun., vol. 5, no. 3, pp. 375–394, Jul. 2021.

[66] K. Gupta, N. Jiwani, and N. Afreen, "A combined approach of sentimental analysis using machine learning techniques," Rev. Intell. Artif., vol. 37, no. 1, pp. 1–6, 2023.