DOI: http://dx.doi.org/10.26483/ijarcs.v15i5.7122

Volume 15, No. 5, September-October 2024

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A HETEROGENEOUS ENSEMBLE MODEL FOR FORECASTING STOCK MARKET MONTHLY DIRECTION

Linus Lavi Raymond
Computer Science Department. Modibbo Adama University
Yola, Nigeria

Etemi Joshua Garba
Computer Science Department. Modibbo Adama University
Yola, Nigeria

Asabe Sandra Ahmadu
Computer Science Department. Modibbo Adama University
Yola, Nigeria

Abstract: It has never been easy to invest in a set of assets, the abnormally of financial market does not allow simple models to forecast future asset values with higher accuracy. Machine learning, which consists of making computers perform tasks that normally requiring human intelligence is currently the dominant trend in scientific research. This article aims to ensemble a model using Support Vector Machine (SVM) and Long-Short Term Memory model (LSTM) to predict the Nigerian Stock Exchange values. The main objective of this paper is to see in which precision a Machine learning algorithm can predict and how much the epochs can improve our model.

Keywords: Ensemble Learning; Support Vector Machine, Long Short-Term Memory; Stock Market; forecasting.

I. INTRODUCTION

Stock markets are volatile and traders need to know about their behavior before making trading decisions. Therefore, stock market prediction is very beneficial for traders to decide whether to purchase, hold, or sell a specific stock at the right time to get enormous profit. For this purpose, they need to buy those stocks whose prices are probable to rise and sell those stocks whose prices are probable to decrease in future. If traders predict stock price trends accurately, they would gain enormous profits. But stock markets are volatile and their correct prediction is difficult for traders because numerous external factors like fundamental factors, technical factors, market sentiment and other unforeseen events influence stock markets

[1] There are two prices that are critical for any investor to know: the current price of the investment he or she owns or plans to own, and its future selling price. Despite this, investors are constantly reviewing past pricing history and using it to influence their future investment decisions. Some investors won't buy a stock or index that has risen too sharply because they assume that prices could further fluctuate, while other investors avoid a falling stock, because they fear that it will continue to deteriorate. Stock market is a well-regulated market established not only to serve as a meeting point for the highly liquid and insolvent (potential) investors, but to also support national economy growth and development. In line with the existing theories, the stock market thrives on information. As well-structured the market is positioned, it is not insulated from or immuned against arrival of information of different kinds.

[2] Predicting the stock market is a complex task that involves a multitude of factors, making it challenging to provide accurate forecasts. Traditional methods include technical analysis, which examines historical price patterns and trading volumes, and fundamental analysis, which assesses a company's financial health and economic indicators. The Technical analysis involves the study of historical price charts and trading volumes to identify trends, patterns, and potential market reversal points. It relies on the belief that historical price movements and patterns can provide insights into future price movements. Fundamental analysis involves evaluating a company's intrinsic value by analyzing financial statements, economic indicators, and other relevant factors. The goal is to determine whether a stock is overvalued or undervalued based on its fundamental characteristics.

[3] Both approaches have their strengths and weaknesses. Technical analysis is criticized for being based on historical data and potentially overlooking fundamental factors. Fundamental analysis, on the other hand, may not capture short-term market sentiment and price movements. Successful investors often use a combination of both methods to make well-informed decisions. For example, they may use technical analysis for short-term timing decisions and fundamental analysis for long-term investment choices. Market conditions, investor sentiment, and unexpected events can influence stock prices, and no method can guarantee accurate predictions. Technical and fundamental analysis are complementary tools that investors use to navigate the complexities of the stock market. A comprehensive approach that considers both historical price movements and a company's underlying financial health can provide a more robust foundation for decision-making.

Machine learning (ML) has gained significant attention in the realm of stock market prediction due to its ability to analyze large datasets, identify patterns, and make predictions based on historical and real-time data

However, it's important to note that the stock market is influenced by numerous unpredictable factors, such as geopolitical events, economic changes, and market sentiment, which can make predictions inherently uncertain. While machine learning models can enhance prediction accuracy by identifying patterns and relationships in data, they are not foolproof. Markets can be irrational, and unexpected events can have a significant impact. Traders and investors should approach predictions with caution and diversify their portfolios to manage risks effectively. Furthermore, staying informed about economic indicators, company news, and global events can help individuals make more informed decisions. It's essential to combine various analytical approaches and continuously adapt strategies based on the dynamic nature of financial markets.

[4] Machine learning models have been used to predict stock market prices. A recent survey of the last decade (2011-2021) on methodologies, recent developments, and future directions in stock market prediction using machine learning techniques is available. The study explains the systematics of machine learning-based approaches for stock market prediction based on the deployment of a generic framework. The study critically analyzed findings retrieved from online digital libraries and databases like ACM digital library and Scopus. The study found that advanced machine learning approaches such as text data analytics and ensemble methods have greatly increased prediction accuracies.

Another study used six machine learning techniques, namely Support Vector Regression (SVR), K-nearest Neighbor (KNN), Decision trees (DTs), Random Forest, Artificial Neural Networks (ANNs), Deep learning technique, to predict the future closing price for five companies that are part of the S&P500 index and the closing price of S&P500 index. A review of the literature on forecasting stock market prices using machine learning and deep learning models is available. The review provides a systematic review, performance analysis, and discussion of the implications of forecasting stock market prices using machine learning and deep learning models.

[5] Ensemble learning is a machine learning technique that combines the predictions of multiple models to improve the accuracy and robustness of the final prediction. It is a meta-algorithm that can be applied to any machine learning algorithm. There are three main types of ensemble learning methods: bagging, stacking, and boosting. Bagging involves training multiple models on different samples of the same dataset and averaging their predictions. Stacking involves training multiple models of different types on the same data and using another model to learn how to best combine the predictions. Boosting involves adding ensemble members sequentially that correct the predictions made by prior models and output a weighted average of the predictions.

Ensemble learning is a technique in machine learning, which can be seen as the process employed in training multiple machines learning models, combining their outputs, thereby treating them as a committee of decision makers. The reason for this is; this committee of individual models should have an overall better accuracy, on the average at least, than any single committee member. Ensemble learning creates a group of models that produces low bias and high variance, and then combines them to produce a new model, which comes with a low bias and a low variance; this would overcome the limitation of high variance in the conventional Multilayer Perceptron backpropagation algorithm, which can be frustrating when preparing a final model for making predictions. Due to the stock market volatility, it requires efficient mechanism to unravel the market mysteries for accurate decision on investment

Ensemble learning is a powerful technique that can help improve the accuracy of machine learning models. It is particularly useful when the individual models have high variance or are prone to overfitting. By combining the predictions of multiple models, ensemble learning can help reduce the variance and improve the generalization performance of the final model.

The purpose of this work is to explore multiple algorithms to predict stock market monthly direction using technical analysis and indicator. Using this data, an in-depth investigation using machine learning techniques will be performed in order to create a model for predicting Nigeria stock market movement. The result of this thesis will show that technical information (indicator) can be used as input to a machine learning classifier to create a prediction model that predicts if the market's movement for the following month is „up‟ or „down‟. The present study adds literature to the existing literatures.

II. LITERATURE REVIEW

A comprehensive literature review on the topic of simulating stock exchange involves examining various research studies, articles, and academic papers that contribute to our understanding of simulation techniques and models applied to stock markets. The review comprises of general discussion on the Stock Exchange, the Nigerian Stock Exchange, the traditional model of Stock price prediction, the Machine Learning Model in Stock price Prediction, and Long Short Term Memory (LSTM), Support Vector Machine (SVM) network models, advantages, and applications. The Section also reviewed the related literature, specifically on the Machine Learning Model and the gap from the literature.

Stock Exchange

A stock exchange is a marketplace where securities such as stocks, bonds, and other financial instruments are traded. It is a platform where buyers and sellers come together to exchange securities. The stock exchange provides a regulated and transparent environment for trading securities. The New York Stock Exchange (NYSE) is the world's largest stock exchange by total market capitalization of its listed companies

There are two types of stock exchanges: physical and virtual. Physical stock exchanges are physical locations where traders meet to buy and sell securities. Virtual stock exchanges, on the other hand, are electronic platforms where traders can buy and sell securities from anywhere in the world

[6] The stock exchange plays a crucial role in the economy by providing liquidity to shareholders and an efficient means of disposing of shares. It also helps companies raise capital by issuing shares to investors in the primary market. Stock exchanges operate under strict regulations to ensure that trading is fair and transparent. They are typically overseen by regulatory bodies that monitor trading activity and enforce rules and regulations. In the United States, the Securities and Exchange Commission (SEC) is responsible for regulating the stock market.

Nigerian Stock Exchange

[7] The Nigerian Stock Exchange (NSE) is the principal securities exchange of Nigeria. It was founded in 1960 and is headquartered in Lagos. The Nigerian Exchange Group (NGX) is a leading integrated market infrastructure in Africa servicing the continent's largest economy. NGX provides a

platform for trading in equities, bonds, exchange-traded funds (ETFs), mutual funds, and other securities. NGX also provides market data, indices, and other services to support the trading activities of market participants You can find live share prices of stocks on the Nigerian Stock Exchange.

The Nigerian Stock Exchange is a key player in the Nigerian economy, providing a platform for companies to raise capital and for investors to invest in these companies. The exchange has a market capitalization of over ₦20 trillion as of December 2023.. The NGX All-Share Index (ASI) is the benchmark index of the Nigerian Stock Exchange, which tracks the performance of all listed equities on the exchange. The NGX ASI has been on an upward trend since the beginning of 2021, reflecting the positive sentiment of investors toward the Nigerian economy.

Nigerian Stock Exchange is a vital component of the Nigerian economy, providing a platform for companies to raise capital and for investors to invest in these companies. NGX provides a platform for trading in equities, bonds, ETFs, mutual funds, and other securities, and also provides market data, indices, and other services to support the trading activities of market participants.

Stock Exchange Prediction Methods

[8] There are several methods for predicting stock prices, including statistical methods and machine learning methods (Bhattacharjee & Bhattacharja, 2019). Some of the statistical methods used for stock price prediction include Simple Moving Average, Weighted Moving Average, Exponential Smoothing, and Naive approach. Machine learning methods such as Linear Regression, Lasso, Ridge, K-Nearest Neighbors, Support Vector Machine, Random Forest, Single Layer Perceptron, Multi-layer Perceptron, and Long Short Term Memory are also used for stock price prediction

A comparative study between different traditional statistical approaches and machine learning techniques was conducted to find the best possible method to predict the closing prices of stocks. The study found that machine learning approaches, especially neural network models, are the most accurate for stock price prediction.

Some of the traditional statistical methods used for predicting a stock using noncomputational statistics include Simple Moving Average, Weighted Moving Average, Exponential Smoothing, and Naive approach. A comparative study between statistical approaches and machine learning approaches has been done in terms of prediction performances and accuracy.

Predicting stock prices with high accuracy is a challenging task due to the dynamic and volatile nature of share prices. There are several factors involved in the prediction, such as physical and psychological factors, rational and irrational behavior, and so on. Therefore, it is advisable to use multiple methods for stock price prediction and to consult with financial experts before making any investment decisions

Machine Learning Approach to Stock Exchange prediction

There are several machine learning methods used for stock price prediction. Some of these methods include Linear Regression, Lasso, Ridge, K-Nearest Neighbors, Support Vector Machine, Random Forest, Single Layer Perceptron, Multi-layer Perceptron, and Long Short Term Memory Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is commonly used for processing and predicting time-series data. LSTM is used for building models to predict the stock prices of companies such as Google. Machine learning is used in many sectors. One of the most popular being stock market prediction itself. Machine learning algorithms are either supervised or unsupervised. In Supervised learning, labelled input data is trained and algorithm is applied. Classification and regression are types of supervised learning. It has a higher controlled environment. Unsupervised learning has unlabeled data but has lower controlled environment. It analyses pattern, correlation or cluster.

Ensemble learning

An ensemble model is a collective outcome of individual models, which are combined in such a way that the model outperforms each individual model in most cases. There are a variety of ensemble techniques. Ensembles fall into two categories as homogeneous ensembles and heterogeneous ensembles based on the learning algorithms used.

Homogeneous ensembles use single base learning algorithm, and popular techniques are bagging, boosting, etc. In contrast, heterogeneous ensemble techniques use different base learning algorithms. For example, popular heterogeneous ensembles include methods such as averaging ensembles, stacking ensembles, weighted ensembles, and blending ensembles.

The averaging ensemble takes the mean value of the prediction values of its individual models. The weighted ensem- ble linearly combines regression outputs of different models, where each model gets a weight based on their performances Here, the best performing individual model gets the most weight, and we gave a lower weight to other two individual models. Still, these two models can overcome the prediction by the best model by combining. For example, assume we assign 3/7 weight for the best model and 2/7 weight for other two models. Then the weak individual models with less weight can collectively overrule the best model.

[9] A stacking ensemble trains several individual models in parallel and combines them. There are two stages in stacking. In the first stage, we train a set of individual models on the raw data and make predictions on that raw data to generate more features. Then, another model at the second stage predicts the final test data by utilizing the generated features. Base models or level-0 learning models are the terms used to refer models trained at the first stage of the stacking. The terms meta-model or a level-1 learning model refer to models that train at the second stage of the stacking process. However, in practice there can be multiple layers of stacking. The blending ensembles are like the stacking ensembles, but blending ensembles use a hold-out data set to train the meta-model. Further, in blending this meta-model is a linear model. With regression, a meta-model can be a linear model, such as linear regression or logistic regression.

[10] Ensemble methods combine multiple models to improve accuracy and generalization. Techniques like Bagging (Bootstrap Aggregating) and Boosting combine predictions from multiple models to reduce bias and variance, leading to more robust crime risk mapping models. These machine-learning techniques offer the potential to capture complex patterns, handle large-scale datasets, and improve predictive accuracy in crime risk mapping. They can incorporate various data sources, identify important features, and provide valuable insights for decision-making and resource allocation in crime prevention strategies. Machine learning techniques serve as valuable tools for confirming the identification of underlying patterns within stock time series data. These techniques offer utility in the evaluation and projection of business performance and similar metrics. Collectively, these comparisons

substantiated the superior performance of the hybrid model in relation to the alternative approaches.

Bagging ensemble learning

Bagging, or Bootstrap Aggregating, is an ensemble learning technique that aims to improve the stability and accuracy of machine learning models by combining the predictions of multiple base models. The primary idea behind bagging is to train each base model on a different subset of the training data, and then aggregate their predictions to create a more robust and generalizable model. Fig. 1 illustrates the component of the Bagging ensemble learning technique.

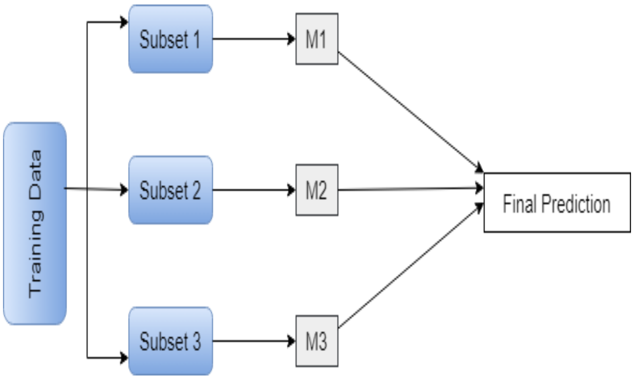

Fig. 1 The architecture of a Bagging Ensemble method

The Fig. 1 describes the component of the model, in which Training data is split into sub-samples which will be fed into the individual models that form the bagging process. The steps involved in the architecture are explained briefly below. Equation 3.1 explains the mathematical process of Bagging.

$$\widehat{f_{bag}} = \widehat{f_1}\,(X) + \widehat{f_2}\,(X) + \cdots + \widehat{f_b}\,(X) \qquad (2.1)$$

the variables are defined as follows:

- $\ddot{f}_{bag}$: The aggregated or bagged prediction. It represents the final prediction obtained by combining the predictions from multiple models.
- $\ddot{f}_1\,(X)$: The prediction made by the ii-th model (where ii ranges from 1 to bb) on the input data X.
- X: The input data or feature set used for making predictions.
- b: The total number of models used in the bagging process.

Long Short-Term Memory (LSTM) Algorithm

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter and Schmidhuber in 1997 and were refined and popularised by many people. They work tremendously well on a large variety of problems, and are now widely used. It basically has three parts to it which are input layer, forget layer, output layer. Input layer is responsible for deciding what amount of information should be carried forward to the next layer from the previous layer and the output layer is responsible for deciding what amount of data should be sent forward into the next layer as input. The reason for the immense popularity of the LSTM is its special power to memorize the data. In a basic neural network that consists of only one layer that is hidden the number of layers to be contained in the input layer mostly depends on the dimensionality of the data, and these input layer neurons get connected to the hidden layers via 'synapses'.

[11] identified the relationship between each of the two nodes from the input layer to the hidden layer consists of a coefficient called weight which acts as a decision maker for the signals. The learning process of the model is basically nothing but a continuous fine-tuning of weights and gradually after the completion of the entire process; the artificial neural networks will have optimum weights for each synapse. An activation function such as a sigmoid or a tangent function

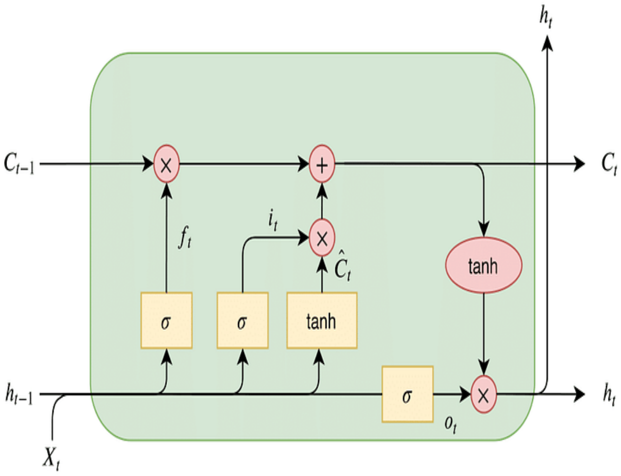

Fig. 2 The architecture of the LSTM Model (Source Man, 2019).

The ability of memorizing sequence of data makes the LSTM a special kind of RNNs. Every LSTM node most be consisting of a set of cells responsible of storing passed data streams, the upper line in each cell links the models as transport line handing over data from the past to the present ones, the independency of cells helps the model dispose filter of add values of a cell to another. In the end the sigmoidal neural network layer composing the gates drive the cell to an optimal value by disposing or letting data pass through. Each sigmoid layer has a binary value (0 or 1) with 0 "let nothing pass through"; and 1 "let everything pass through." The goal here is to control the state of each cell, the gates are controlled as follow: - Forget Gate outputs a number between 0 and 1, where 1 illustration "completely keep this"; whereas, 0 indicates "completely ignore this." - Memory Gate chooses which new data will be stored in the cell. First, a sigmoid layer "input door layer" chooses which values will be changed. Next, a *tanh* layer makes a vector of new candidate values that could be added to the state. - Output Gate decides what will be the output of each cell. The output value will be based on the cell state along with the filtered and freshest added data.

Applications of LSTMs

[13] Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is commonly used for processing and predicting time-series data. LSTM is used for building models to predict the stock prices of companies such as Google. LSTM applies to the classification, processing, and prediction of data based on time series, such as in handwriting, speech recognition, machine translation, speech activity detection, robot control, video games, and health.
Some of real life applications of LSTM as folllows:
i. Univariate forecasting: the most common and most obvious way to use the lstm model is when doing a simple univariate

forecasting problem. although the model fits many parameters that should make it sophisticated enough to learn trends, seasonality, and short-term dynamics in any given time series effectively, i have found that it does much better with stationary data (data that doesn't exhibit trends or seasonality). so, with the air passengers dataset — which is available on kaggle with an open database license — we can easily create an accurate and reliable forecast using fairly simple hyperparameters, if we simply detrend and de-season the data.

ii. Multivariate forecasting: let's say that we have two series that we expect move together. we can create a lstm model that takes both series into consideration when making predictions with the hope that we improve the model's overall accuracy. this is, of course, multivariate forecasting.

iii. Probabilistic forecasting: probabilistic forecasting refers to the ability of a model to not only make point predictions, but to provide estimates of how far off in either direction the predictions are likely to be. Probabilistic forecasting is akin to forecasting with confidence intervals, a concept that has been around for a long time. A quickly emerging way to produce probabilistic forecasts is by applying a conformal confidence interval to the model, using a calibration set to determine the likely dispersion of the actual future points. This approach has the advantage of being applicable to any machine learning model, regardless of any assumptions that model makes about the distribution of its inputs or residuals. It also provides certain coverage guarantees that are extremely useful to any ml practitioner. We can apply the conformal confidence interval to the lstm model to produce probabilistic forecasts.

Iv. Dynamic probabilistic forecasting: the previous example provided a static probabilistic prediction, where each upper and lower bound along the forecast is equally far away from the point estimate as any other upper and lower bound attached to any other point. When predicting the future, it is intuitive that the further out one attempts to forecast, the wider the error will disperse — a nuance not captured with the static interval. There is a way to achieve a more dynamic probabilistic forecast with the lstm model by using back testing.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a training algorithm for learning classification and regression rules from given data. A kernel-based method, which can be used with linear, polynomial, radial basis function (RBF) and other custom kernel functions. Structural Risk Minimization (SRM) principle to develop binary classifications. Since in this study three data labels are used and SVM is a binary classification tool, meaning that it accepts only two classes at a time, an approach is adopted to deal with the three classes. The SVM implementation used is brought from the publicly available LIBSVM, V3.17. The approach used is one-against-one classification where SVM trains each class against another, having the most common prediction determines the output data label.

The multiple green lines represent good solutions to the problem, however, the bold yellow line shows the optimal hyperplane which maximizes the separation between the two classes. In SVM, a hyperplane is considered incorrect if it passes close to the training points as it will be noise sensitive and it will not generalize correctly. Thus, the goal is to find the best separating hyperplane which maximize the margin from all training data.

[14] the SVM classification model accepts training data, training labels and a variable which contains information about how to train the SVM classifier to generate a model. The choice of the SVM training data, kernel and kernel parameters are automated by a cross validation-based model. In the cross-validation, the SVM algorithm is trained with both normalized data as well as not normalized data. The SVM algorithm is trained with four types of kernels, namely: linear, polynomial, radial basis and sigmoid kernel. Another parameter is the cost parameter, which is investigated within a range of 10-6 to 103 in sequence of one.

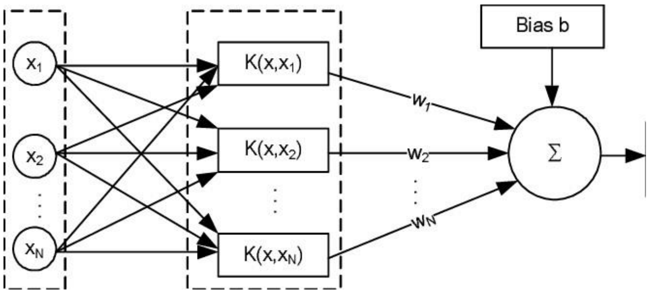

Fig. 3 The architecture of the Support Vector Machine (Vapnik 1995)

SVM Parameters

The tuning parameters of SVM classifier are kernel parameter, gamma parameter and regularization parameter.

Applications of SVM

SVMs find application in various fields like Handwritten digit and character recognition,

Object detection and recognition, Speaker identification, Benchmarking time, Series prediction tests, Text classification, Biometrics, Content-based image retrieval, Image classification.

How SV M works

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, and then the data are transformed in such a way that the separator could be drawn as a hyperplane.

There are two general classes of machine learning techniques. The first is supervised learning, in which the training data is the target where each example is collections of features that determine with the correct output corresponding to the feature set. This means that the algorithm is given features and outputs for a particular dataset (training data), and must apply what it "learns" from this dataset to predict the outputs (target) for another dataset (test data). Unsupervised learning, on the other hand, consists of examples where the feature set is untargeted.

[14] Supervised learning can be further broken down into classification and degeneration problems. In classification problems there are a set number of outputs that a feature set can be target as, whereas the output can take on continuous values in degeneration problems. In this I am trying to treat the problem of close stock price market forecasting as a classification problem The feature set of a stock‟s recent price volatility and momentum, along with the index's recent volatility and momentum, are used to predict whether

or not the close stock's price in this month or in the future will be higher (+1) or lower (-1) than the current day's price.

III. METHODOLOGY

To build our model we are going to use the ensemble LSTM and SVM, our model will be structured as follows:

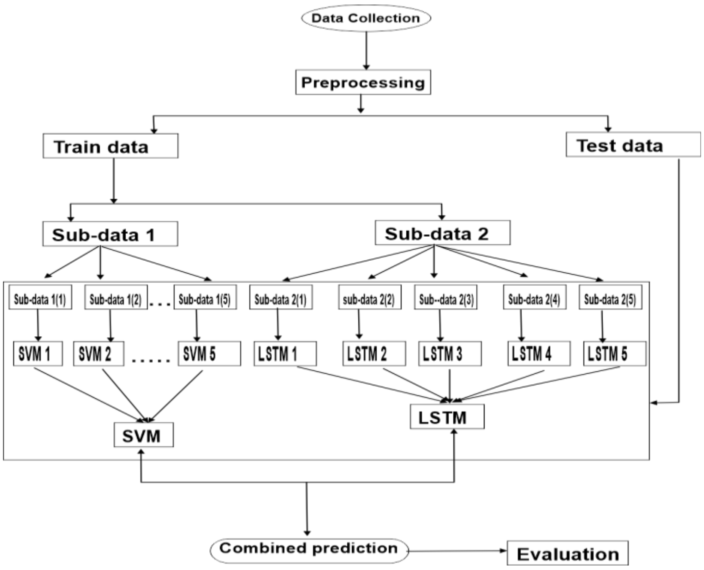

Fig. 4 Sketch of the proposed research methodology.

The overall flow of the proposed ensemble LSTM and SVM is demonstrated in Figure 3.4. The above diagram is a complete architecture of the system that illustrates the data flow from the source, to be processed, split into Train and Test sets, then the trained set split into two for the two heterogeneous ensembled models, then to be passed to the model for proper process. After the processes the test model is used for the outcome and a combined aggregation of the model output is formed, which will be evaluated.

Data Collection

Data collection is a crucial step in building machine-learning models for predicting stock exchange movements. The quality and relevance of the data directly impact the model's accuracy and effectiveness. Understanding your goals will guide the selection of relevant data. The main objective of this study is to predict stock prices, and market direction (e.g., up, down, or stable). To justify the model, this study uses a similar dataset to the existing one i.e. the Nigerian Stock Exchange dataset. The data can be searched and downloaded in .csv (comma-separated value) format between two dates using the stock ticker symbol of a company or stock market.

The data were collected by using secondary method of data collection from the index of the Nigerian Stock Exchange, Lagos. Historical data were downloaded from the Nigerian section of 'investing.com', a credible stock market analysis website. The data collected is for a period of 252 days starting from July 2, 2020 to July 2, 2021 excluding weekends and public holidays.

Data acquisition and description

The data for training and testing the model will be collected from Nigerian stock exchange daily price lists. It is the stock price data for NSE. The data is a daily trading and it comprised of the opening price, closing price, high price, low price, adjacent close, and volume of each trading day. The data to be considered or taken into account will be readily available in the csv format which will first be read and then be converted into a data frame by making use of one of the most popular libraries, Pandas in Python. Although Machine learning has various algorithms that could be used for predicting the stock prices here in this paper, we will make use of two main algorithms known as an ensemble LSTM with SVM.

Population size of the data

The data was obtained from Nigerian stock exchange daily trading price list and utilised in this experiment. The data set consists of 252 features spanning from 2020-07-02 to 2021-08-02 and comprises seven columns labeled Date, Open, High, Low, Close, Adj Close, and Volume. These features express respectively the stock traded date, opening price, highest selling price, lowest selling price, closing price, the number of shares traded, and the closing price when dividends are paid to investors.

The data can be searched and downloaded in .csv (comma-separated value) format between two dates using the stock ticker symbol of a company or stock market.

The data were collected by using secondary method of data collection from the index of the Nigerian Stock Exchange, Lagos. Historical data were downloaded from the Nigerian section of 'investing.com', a credible stock market analysis website. The data collected is for a period of 252 days starting from July 2, 2020 to July 2, 2021 excluding weekends and public holidays.

IV. RESULTS AND DISCUSSION

This chapter describes the proposed system's general implementation processes, result analysis, and interpretation. The implementation stages followed the proposed study objectives. Hence the implementation flows from data collection and preprocessing, using lag feature engineering techniques for fast and effective time series data to ensure that the models capture relevant patterns and trends. Developed Nigerian Stock Exchange prediction model using an Ensemble of five ensembled Long Short-Term Memory (LSTM) networks and five ensembled Support Vector Machine (SVM). Optimize the LSTM network hyperparameters to achieve the best possible predictive performance for improving prediction with higher accuracy. Finally to Examine and compare execution times, accuracy, and error metrics of techniques in (iii) over-stock data from NSE.

Data Collection and Preprocessing

The dataset utilized for training and testing the model originates from the Nigerian Stock Exchange (NSE) daily price lists, spanning the period from 2020 to 2021. It encompasses crucial elements of daily trading activity, including the opening price, closing price, high price, low price, adjusted close, and volume for each trading day. The opening price marks the initial valuation at the start of trading, while the closing price signifies the day's final valuation, both influencing trading decisions. High and low prices denote the peak and nadir of stock valuation during the day, respectively, offering insights into price volatility and support levels. The adjusted close price accounts for corporate actions, ensuring accurate performance assessment. Finally, the volume reflects market activity and liquidity, aiding in understanding investor interest and potential price movements. This comprehensive dataset empowers analysts and machine learning practitioners to develop models for predicting stock price movements and making informed investment decisions, leveraging the richness of its components for robust analysis. The same data was used for the existing system modeling. This is employed to fairly the two model performances. Figure 4.1 is a snapshot of the proposed dataset.[15] Obtained their data from Nigerian stock exchange daily trading price list and utilized in this experiment. The data set consists of 252 features spanning from 2020-07-02 to 2021-08-27 and comprises eight columns labeled Serial number, Date, Open, High, Low, Close, Adj Close, and Volume. These features express respectively the stock traded date, opening price, highest selling price, lowest selling price, closing price, the number of shares traded, and the closing price when dividends are paid to investors.

The eight features are depicted in the chart for more comprehension to ascertain their impact. The distribution of open prices in a stock exchange reflects how these prices are spread out. In a healthy exchange, they often approximate a normal distribution, with most prices clustering around the average. Factors like market volatility, trading volume, and economic conditions influence this distribution. Events such as earnings announcements or geopolitical developments can cause sudden shifts or outliers. Understanding this distribution is vital for investors and analysts to gauge market dynamics and make informed decisions. Figure 4.2 illustrates the assertion.

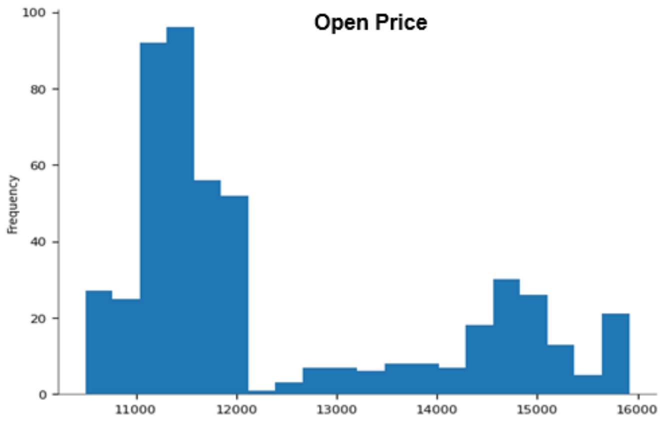

Fig. 5 The Distribution of Open Price of the Nigeria Stock Exchange.

The distribution of high prices in the stock exchange reflects how prices are spread across different stocks. It's influenced by factors like market demand, company performance, and investor sentiment. While a healthy market might show a normal distribution, events like earnings reports or geopolitical tensions can cause deviations. Understanding this distribution helps investors gauge market trends and manage risk. Figure 4.3 depicts the Distribution of High prices of the Nigeria Stock Exchange

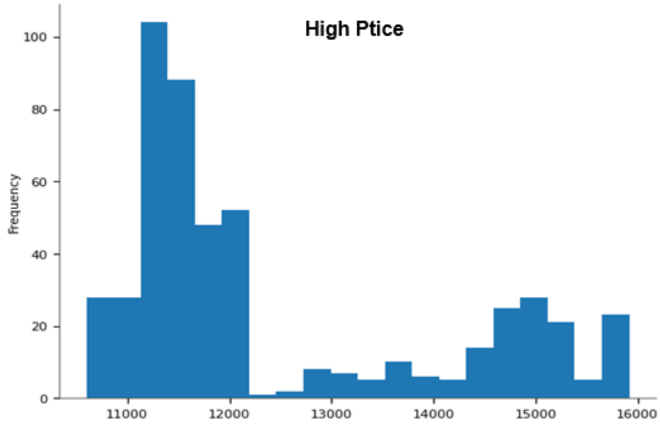

Fig. 6 The Distribution of High Prices of the Nigeria Stock Exchange

The distribution of low prices in the stock exchange illustrates how prices are dispersed among various stocks. Market dynamics, company fundamentals, and investor behavior shape it. Generally, a normal distribution might indicate a stable market, but external events can cause fluctuations. Analyzing this distribution assists investors in identifying trends and assessing risk levels. Figure 4.3 shows how the distribution of low prices in the Nigeria stock exchange

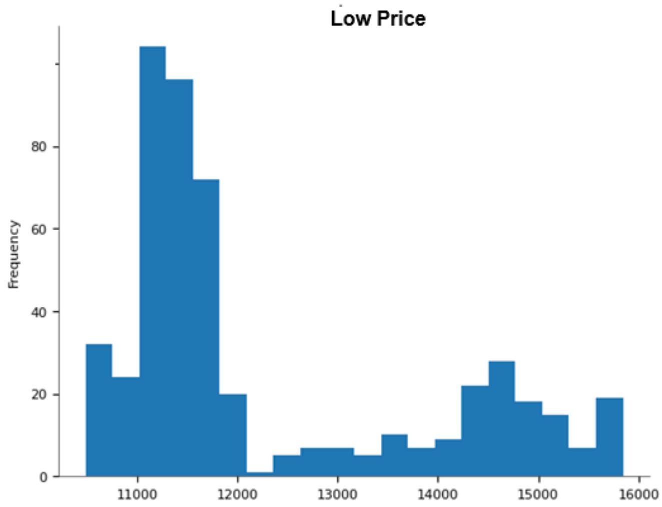

Fig. 7 The Distribution of Low Prices of the Nigeria Stock Exchange

The distribution of closing prices in the stock exchange reflects how prices are spread across different stocks at the end of trading sessions. This distribution is influenced by factors such as market sentiment, economic indicators, and company performance. Understanding this distribution helps investors gauge market trends, identify potential investment opportunities, and manage risks associated with their portfolios. Figure 4.5 shows the Close prices of the Nigeria Stock Exchange

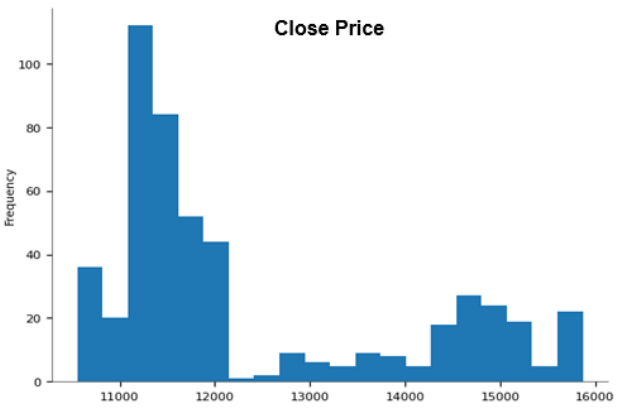

Fig. 8 The Distribution of Closing Prices of the Nigeria Stock Exchange

Fig. 8 illustrates the relationship of Dates with the five most important features in the Nigeria Stock Exchange dataset lists trading dates in "MM/DD/YYYY" format, organizing data chronologically. This organization enables trend analys pattern recognition, and correlation assessments with exter factors, aiding analysts in evaluating stock performance a making informed investment decisions. Figure 4.6 illustra the time series relationship of the Date with the Open price the Nigeria Stock Exchange.

The relationship between the date and the opening price of t stock market index appears to fluctuate over time. From Ju 2020 to September 2020, there is a general upward trend in b variables, indicating a potential positive correlation. Howev from September to December 2020, there seems to be m volatility in the opening prices despite some fluctuations in t date. In the early months of 2021, the relationship seems l discernible, with periods of fluctuation and stability in both t date and the opening price. Overall, while there may be so correlation between the date and the opening price duri certain periods, the relationship is likely influenced by vari factors such as market sentiment, economic conditions, a geopolitical events. Figure 4.6 (a) depicts the Time ser relationship between date and Open and Close prices for t Nigeria Stock Exchange.

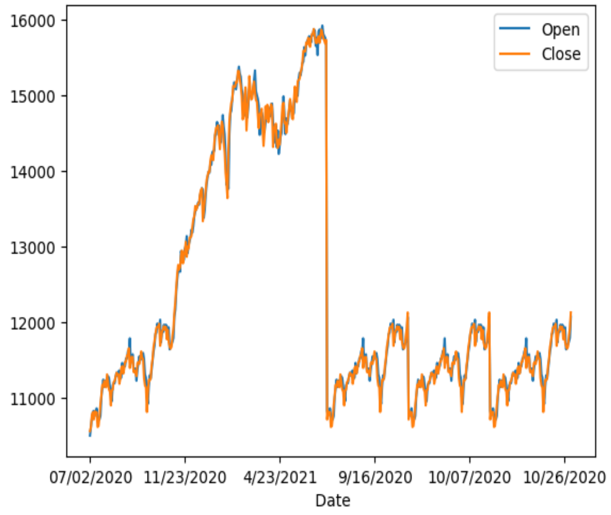

Fig. 9 The Time series relationship of the date and Open and Close prices for the NSE

The relationship between the date and close price in the Nigerian stock market typically demonstrates the historical performance and trends of a particular stock or index over time. Analyzing this relationship allows investors to identify patterns, such as seasonal fluctuations or long-term trends, which can inform their investment decisions. Factors such as economic conditions, company earnings reports, geopolitical events, and market sentiment all contribute to the fluctuations in close prices over time. By studying this relationship, investors can gain insights into market behavior, assess the performance of their investments, and develop strategies to capitalize on potential opportunities or mitigate risks.

The relationship between the date and high price in the stock market reflects market sentiment, economic conditions, and investor behavior. High prices are influenced by factors like positive earnings reports, economic indicators, corporate developments, and external events. Analyzing this relationship helps investors identify trends, gauge market volatility, and make informed decisions. Figure 4.6 (b) depicts the Time series relationship between the date and High prices for the Nigeria Stock Exchange

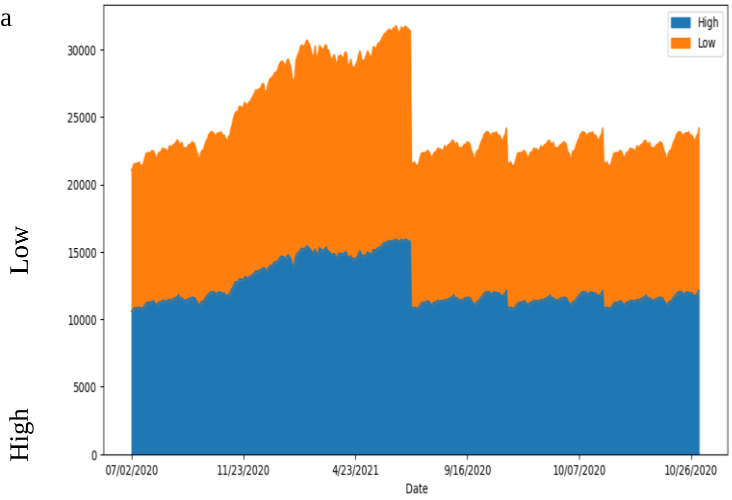

Fig. 10 The Time series relationship of the date and High and Low prices for the NSE

The relationship between the date and low price in the stock market can reveal patterns of market volatility, investor sentiment, and economic conditions. Low prices are often influenced by factors such as negative news, economic uncertainties, geopolitical events, and profit-taking by investors. Understanding this relationship allows investors to assess risk levels, identify potential buying opportunities, and adjust their investment strategies accordingly.

The relationship between the date and trading volume in the stock market indicates market activity and investor participation over time. High trading volumes on certain dates may suggest increased investor interest or significant news events influencing market sentiment. Conversely, low trading volumes may indicate investor caution or a lack of significant market-moving news. Analyzing this relationship helps investors understand market dynamics, identify potential trend reversals, and assess the overall market sentiment. Figure 4.6 (c) Shows the Time series relationship of the date and volume for the Nigeria Stock Exchange

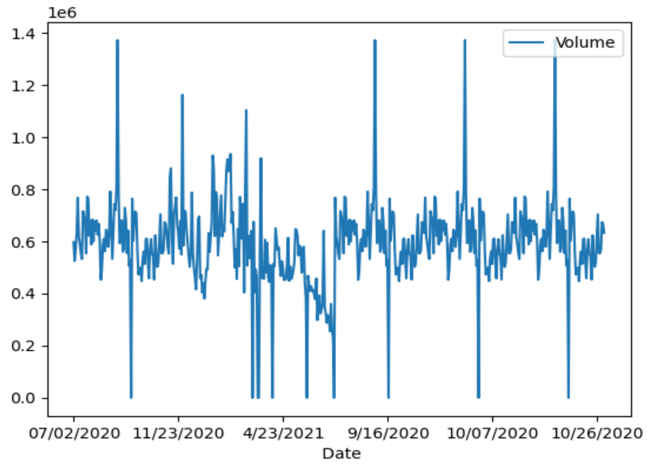

Fig. 11 The Time series relationship of the date and Volume for the NSE

After scientific processes of preprocessing the most relevant features considered for the modeling were; Date Open and Close prices high and Low prices and Volume.

Implementing Lag Feature Engineering

Lag feature engineering, also known as lagging or time lagging, is a technique used in time series analysis and forecasting to create new features based on past values of existing features. It involves shifting the values of a feature backward in time by a certain number of periods, creating a lagged version of the feature.

[16] Lag features to a data frame for each specified feature. It iterates over each feature in the list `features` and for each feature, it creates lagged versions up to a specified lag window (`lag`). For example, if `lag` is set to 3, it creates lagged versions of each feature for 1, 2, and 3 time periods ago. These lagged features are appended to the DataFrame with column names indicating the original feature name followed by **"_lag_"** and the lag period (e.g., "feature_lag_1", "feature_lag_2", etc.). After adding the lagged features, the code drops rows with NaN (Not a Number) values, which occur because the lagged features cannot be calculated for the first few rows of the DataFrame. This ensures that the DataFrame only contains complete data with lagged features.

The lag technique was applied to the dataframe of the proposed Dataset features before passing the dataframe to the two ensemble models the Ensemble SVM and Ensemble LSTM. The lag technique, or lag feature engineering, involves incorporating past values of features into a dataset to capture temporal patterns and dependencies, especially in this model of time series analysis. By adding lagged features, the model understands how historical data influences current and future behavior, potentially improving predictive accuracy and stabilizing data stationarity.

Ensemble of five ensembled Long Short-Term Memory (LSTM) networks

The code snippet in Appendix iv performs time series forecasting using an ensemble of LSTM (Long Short-Term Memory) neural networks. It starts by adding lagged features to the dataset to capture temporal dependencies. Then, it splits the data into train and test sets, normalizes the input features, and reshapes them to fit the LSTM model. Next, it defines an ensemble of five LSTM models, each with a similar architecture. The models are trained on the training data and used to make predictions on the test data. The predictions from all models are averaged to obtain the final ensemble prediction. Finally, the root mean squared error (RMSE) is calculated to evaluate the performance of the ensemble model.

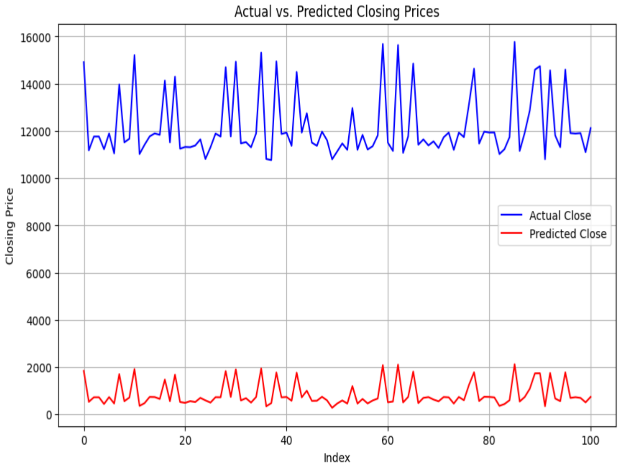

Fig. 12 LSTM Model prediction model of Actual vs Predicted Closing price

The ensemble root mean squared error (RMSE) of approximately 11480.51 indicates that the ensemble model, which comprises five LSTM models, produces predictions that deviate, on average, by around $11480.51 from the actual closing prices in the test dataset. This level of error could be considered relatively high in the context of financial forecasting, where precision is often crucial. Such a high RMSE suggests that the model may not be effectively capturing all the relevant patterns and nuances present in the data, potentially leading to suboptimal trading decisions or inaccurate risk assessments. Consequently, further analysis and refinement of the model architecture, feature engineering techniques, or dataset may be necessary to improve predictive accuracy and reduce the error to a more acceptable level for practical use in financial markets. Additionally, alternative modeling approaches or ensemble strategies could be explored to enhance the model's performance and robustness in predicting stock prices.

Ensemble of five ensembled Support Vector Machines(SVM)

The provided code in Appendix V loads a dataset containing stock data and selects several input variables such as Open, High, Low, Close, Adj Close, and Volume. Lag features are created for each input variable, ranging from lag 1 to lag 5, capturing historical data points. These lag features are combined with the target variable (Close) and the dataset is split into training and testing sets. Support Vector Machine (SVM) models are trained individually on the training data, resulting in five SVM models. Then, a Bagging Regressor ensemble model is trained on the same training data. The code evaluates each SVM model and the Bagging Regressor model using mean squared error (MSE) on the test set. The resulting MSE values provide insights into the performance of each SVM model and the ensemble Bagging Regressor model. To visualize the performance, we can create a bar chart showing the MSE values of each SVM model and the Bagging Regressor model.

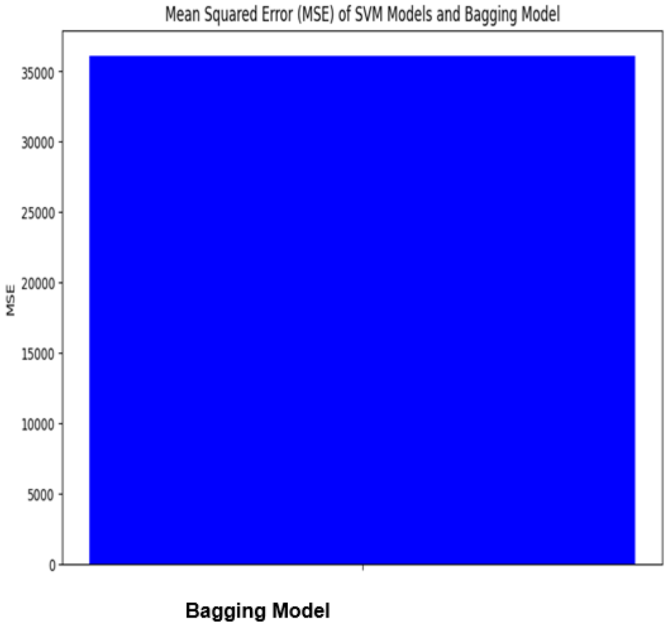

Fig. 13 MeanSquared Error of an Ensemble SVM Prediction Model

The Bagged Ensemble RMSE of 6057.90 signifies the overall predictive accuracy of the ensemble model formed through Bagging Regression. Figure 4.8 shows the Mean Squared Error of an Ensemble SVM Prediction Model. This metric, Root Mean Square Error (RMSE), gauges the average disparity between predicted and actual values. A lower RMSE indicates greater accuracy in predicting the target variable (Close) than individual SVM models. Consequently, the ensemble method effectively harnesses the diversity among SVM models to minimize prediction errors, enhancing the robustness and reliability of stock price predictions.

Aggregating the two Ensemble Models

The Bagged Ensemble RMSE (Aggregate) of 11480.508139576461 represents the collective predictive performance of an ensemble formed by aggregating predictions from LSTM and SVM models. This metric measures the average disparity between the predicted and actual values of the target variable (Close). The higher RMSE compared to individual models suggests that the aggregate approach may not provide significant improvement in predictive accuracy over using LSTM or SVM models individually. However, further analysis is necessary to explore potential synergies between these models and optimize the ensemble's performance.

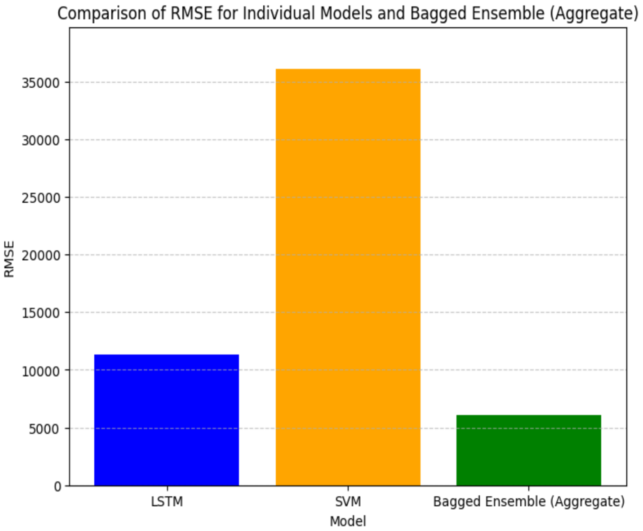

Fig. 14 Comparing RMSE for individual Models and Bagged Ensemble Model

The results indicate that the Bagged Ensemble model has the lowest RMSE (Root Mean Squared Error) of 6057.90, followed by the LSTM Ensemble with an RMSE of 11372.11, and the SVM Bagging model with an RMSE of 36124.36. In terms of accuracy, lower RMSE values indicate better performance, suggesting that the Bagged Ensemble model outperforms both the LSTM Ensemble and the SVM Bagging model in predicting the target variable. The substantial difference between the SVM Bagging model and the other two models suggests that the SVM-based approach might not be as effective in this scenario compared to the LSTM-based and Bagged Ensemble approaches.

Examine and compare execution times of the models

Comparing the execution types of the three ensembles, the LSTM Ensemble had the longest execution time, taking approximately 1.28 seconds to complete. Despite this longer time, it achieved an RMSE of 11496.03. On the other hand, the SVM ensemble executed much faster, only requiring around 0.046 seconds, yet it obtained a slightly higher RMSE of 1400.58. The combined ensemble, however, exhibited the shortest execution time of merely 0.001 seconds while achieving an RMSE of 6106.67, striking a balance between speed and accuracy. Therefore, if prioritizing speed, the SVM ensemble is preferable, but for a better compromise between speed and accuracy, the combined ensemble serves as a viable option. Figure 4.10 depicts the execution time of the three ensemble models.

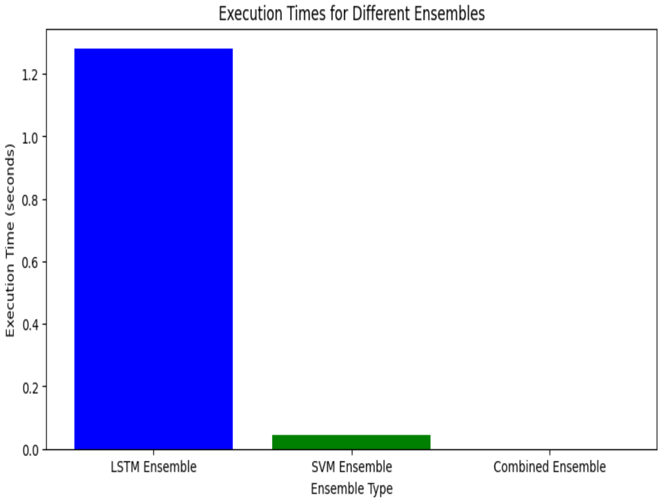

Fig. 15 Execution time of the three Ensemble Model

Result Discussion

The proposed improved prediction model combines the strengths of Ensemble Long Short Term Memory (LSTM) and Support Vector Machine (SVM) techniques to forecast the monthly direction of the Nigeria Stock Exchange. The Bagged ensemble model, which aggregates predictions from both LSTM and SVM models, yielded an RMSE of 11480.51, indicating its predictive performance. While this RMSE is higher than that of individual models, it signifies a collective effort to enhance accuracy. Comparatively, the Bagged ensemble outperforms the LSTM ensemble and the SVM Bagging model in terms of RMSE, suggesting superior predictive capability. Moreover, analyzing execution times reveals that the combined ensemble strikes a balance between speed and accuracy, making it a promising choice for practical applications. Overall, the integration of LSTM and SVM models in the ensemble approach presents a compelling solution for forecasting stock market direction in Nigeria, offering improved accuracy while maintaining efficiency.

V. CONCLUSION

The objective is to use an ensemble LSTM and SVM machine learning techniques to determining whether the overall stock market index is forecast to rise or fall in the coming month which will help investors to make more informed and accurate investment decisions. We propose a stock price prediction system that ensemble deep learning, machine learning, and other external factors for the purpose of achieving better stock prediction accuracy and issuing profitable trades. The purpose behind this survey is to classifying the current techniques related to adapted methodologies, used various datasets, performance matrices, and applying techniques. The techniques used in the stock market prediction are categorized in different ML algorithms. For improving the prediction accuracy, some of the selected studies use the hybrid approaches in the stock market. LSTM and SVM techniques are widely used approach for achieving the successful stock market prediction. The biggest challenge the stock market prediction face is that most current techniques cannot be identified with the aid of historical data on stocks. Hence stock markets are influenced by other factors such as policy decisions by government and consumer sentiments. LSTMs are best used for predicting numerical stock market index values. Support vector machines best fit classification problems. In the future, we will strive to improve the system for making a reliable stock market system that is more reliable and accurate in order to save investors from uncertainty when making investment decisions.

VI. ACKNOWLEDGMENTS

First and foremost I would like to acknowledge God for whom all things are possible. All the glory and honor belongs to him.

I would like to pay my heartfelt respect and gratefulness to my supervisor Professor E.J. Garba, for his excellent guidance, and also for his efforts to correct my thesis draft and revising technical terms. I am extremely thankful for his inspiring ideas that pointed me in the right direction. His patience, enthusiasm, superior knowledge and stimulating suggestions has really helped me and encouraged me. Also to my Co-Supervisor Dr. A.S. Ahmadu, she's not only my supervisor but a great teacher, a wonderful mentor and mother as well.

I want to thank in a special way the peer- editors of this work for taking their time to read through and to correct all my errors and providing valuable suggestions that further improved my work, may God reward your efforts. I acknowledged that errors or omissions, if any, are my own responsibility.

REFERENCES

[1] R. Lekhani, "Stock Prediction using Support Vector Regression and Neural Networks". *International Journal of Advance Research*, *Ideas and Innovations in Technology*, vol. *3*, 2017.

[2] I. K.. Nti, A. F. Adekoya, and B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions". *Artificial Intelligence Review*, vol. *53*, 2020, pp 3007-3057.

[3] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting:". A literature review. *Expert Systems with Applications*, vol. *197*, pp. 116659.

[4] N. Rouf, M. B. Malik, T. Arif, S. Sharma, S. Singh, S. Aich, and H. C. Kim, "Stock market prediction using machine learning techniques" A decade survey on methodologies, recent developments, and future directions. *Electronics*, vol. *10*, 2021, pp. 2717.

[5] A. F. Kamara, E. Chen, and Z. Pan, "An ensemble of a boosted hybrid of deep learning models and technical analysis for forecasting stock prices". *Information Sciences*, vol. *594*, 2022, pp. 1-19.

[6] F. Bio, "Machine learning and its applications to biology". *PLoS computational biology*, *3*6, 2022, pp. 116

[7] M.H. Kwayist, "*Nigerian Stock Exchange (NGX) Live.* African x changes *https://afx.kwayisi.org/ngx/*

[8] I. Bhattacharjee, and P. Bhattacharja, "Stock price prediction: a comparative study between traditional statistical approach and machine learning approach". In *2019 4th international conference on electrical information and communication technology (EICT),* pp. 1-6.

[9] A. P. Wheeler, and W. Steenbeek, "Mapping the risk terrain for crime using machine learning". *Journal of Quantitative Criminology*, vol. *37*, 2021, pp. 445-480.

[10] A. Namdari, and Z. S. Li, "Integrating Fundamental and Technical Analysis of Stock *International". Journal of Computer Applications, vol. 128,* 2018, *pp. 18–21.*

[11] A. Biswal, "Bagging in Machine Learning: A review. Progress in Advanced Computing and Intelligent Engineering". Proceedings of ICACIE 2020, 2021, pp. 323-333.

[12] P. Chhajer, M. Shah, and A. Kshirsagar, "The applications of artificial neural networks, support vector machines, and long–

short term memory for stock market prediction”. *Decision Analytics Journal*, vol. *2*, 2022, pp.100015

[13] A. Aaryan, B. Kanisha, “Forecasting stock market price using LSTM-RNN”. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) 23, 2022, pp. 1557-1560.

[14] K. H. Sadia, A. Sharma, A. Paul, S. Padhi, and S. Sanyal, .Stock market prediction using machine learning algorithms”. *Int. J. Eng. Adv. Technol*, *8*, 2019, pp. 25-31.

[15] L.L. Raymond, E.J. Garba, and Y.M. Malgwi, “Application of Support Vector Machine in Predicting Stock Market Monthly Direction”. *International journal of Advances in Engineering and Management, Vol. 3,* 2021, pp. *1-12*

[16] E. Lucena-Sánchez, F. Jiménez, G. Sciavicco, and J. Kamińska, “Simple versus composed temporal lag regression with feature selection, with an application to air quality modeling”. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems,* IEEE, 2020, pp. 1-8.