International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 31 Issue 01 – JANUARY 2025.

Ensemble of AI Models and Predictive Analytics for Stock Market Monthly Direction

Lavi Linus Raymond*, Etemi Joshua Garba*, Asabe Sandra Ahmadu*

*Computer Science Department. Modibbo Adama University Yola, Nigeria.

***Abstract*— 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 effectively.**

***Index Terms*—Ensemble Learning, Long-Short Term Memory, Support Vector Machine, Stock Market, Prediction.**

I. INTRODUCTION

[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.

[2]. 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.

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.

[3]. 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.

[4]. 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.

[5]. 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. In the work of Sonkavde, 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.

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.

[6]. 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.

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

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.

A. Stock Exchange

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.

B. 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.

C. Stock Exchange Prediction Methods

There are several methods for predicting stock prices, including statistical methods and machine learning methods. 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.

[8] 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.

D. Statistical methods of stock exchange prediction

[9]. There are several statistical methods used for stock price prediction. Some of these methods include Simple Moving Average, Weighted Moving Average, Exponential Smoothing, and Naive approach. These methods are based on the assumption that future stock prices will follow the same pattern as past prices.

E. Simple Moving Average (SMA)

A Simple Moving Average (SMA) is a statistical method that calculates the average of a stock's price over a specified time. The SMA is calculated by adding up the closing prices of a stock over a specified period and then dividing the sum by the number of periods[1]. The SMA is used to smooth out the fluctuations in a stock's price and to identify trends. prices. The SMA is commonly used in technical analysis to identify support and resistance levels[1]. The SMA is also used in conjunction with other technical indicators to make investment decisions.

F. Weighted Moving Average (WMA)

The Weighted Moving Average (WMA) is a statistical method that is similar to the Simple Moving Average (SMA), but it assigns more weight to the most recent prices. The WMA is calculated by multiplying each price in the data set by a weight factor and then dividing the sum of the products by the sum of the weights . The weight factor is determined by the number of periods in the moving average. The most recent prices are assigned a higher weight factor, while the older prices are assigned a lower weight factor.

The WMA is used to smooth out the fluctuations in a stock's price and to identify trends[2]. The WMA is a lagging indicator, which means that it is based on past prices and does not predict future prices. The WMA is commonly used in technical analysis to identify support and resistance levels. The WMA is also used in conjunction with other technical

indicators to make investment decisions.

G. Naive Approach

The Naive approach is a simple statistical method used for stock price prediction. The method assumes that the future stock price will be the same as the current price. The method is based on the assumption that future stock prices will follow the same pattern as the past price. The Naive approach is used to smooth out the fluctuations in a stock's price and to identify trends. The method is a lagging indicator, which means that it is based on past prices and does not predict future prices. The Naive approach is commonly used in technical analysis to identify support and resistance levels. The Naive approach is also used in conjunction with other technical indicators to make investment decisions.

H. The need for stock market forecasting

[10]. In the stock market, the investor shows interest in profit by investing some money in the stock market. The stock market has shown investor interest due to advanced applications where prediction may lead to prosperous market forecasting. Predicting movements of the stock market precisely depends on advance information. The tools which are used for stock market forecasting can track and control the market which can be used to make the right decisions. The stock market needs to handle several information on industrial stocks which covers the entire financial.

These are adjusted according to the business status investors who consider sales and purchase. Several factors affect the market position are the future estimation income, a news release on earnings and changes in management, etc. Therefore, accurate prediction of the stock market helps investors in making better decisions. Through ML techniques the investor can earn more money with high risk.

Machine Learning Approach to Stock Exchange prediction

[11]. 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.

I. Ensemble learning

[12]. 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.

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.

J. 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. Figure 3.2 illustrates the component of the Bagging ensemble learning technique.

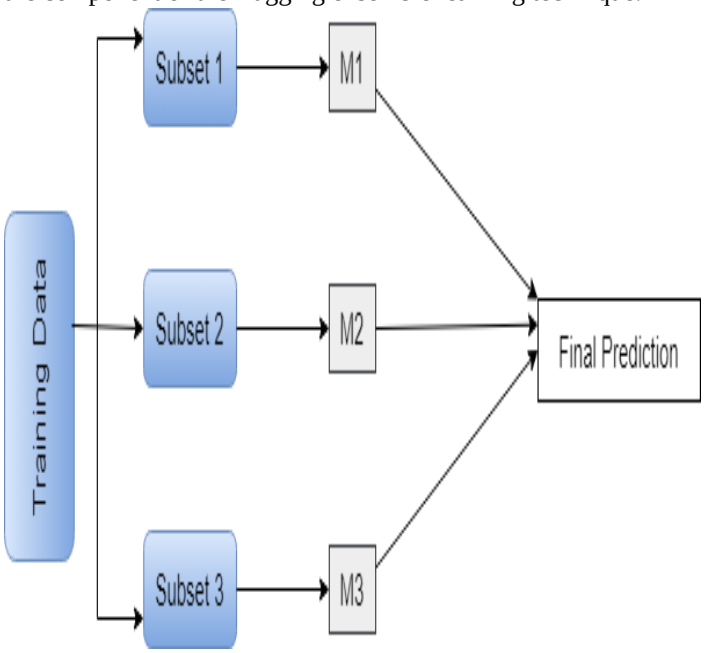

Figure 2.1 The architecture of a Bagging Ensemble method

The Figure 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.

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

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.

1) Steps in Bagging

a) Data Sampling

b) Randomly sample subsets (with replacement) from the training data to create multiple bootstrap samples. Each subset is used to train a separate base model.

2) Base Model Training

Train a base model on each bootstrap sample. These models can be trained using the same algorithm or different algorithms.

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'.

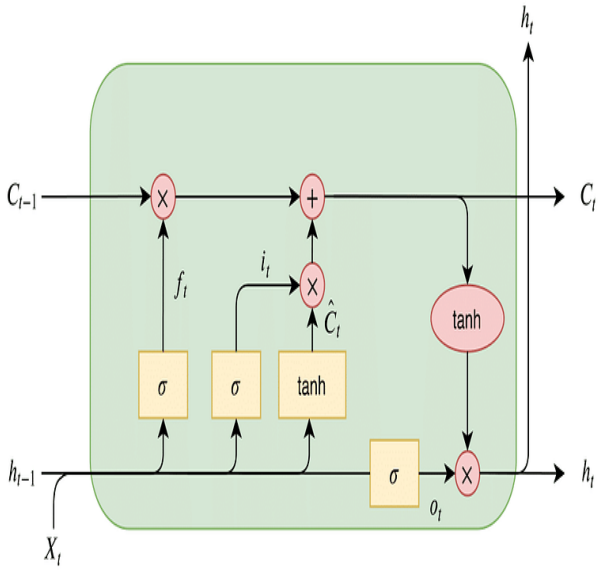

Figure 2.2 The architecture of the LSTM Model

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.

K. The Component of the Architecture

The System that we propose has several steps that lead to model generation. The steps involved are as follows.

1) Stage 1: Raw Data: In this stage, the historical stock data is collected from the internet and this historical data is used for the

prediction of future stock prices.

2) Stage 2: Data Pre-processing: The pre-processing stage involves

a) Data discretization: Part of data reduction but with particular importance, especially for numerical data

b) Data transformation: Normalization.

c) Data cleaning: Fill in missing values.

d) Data integration: Integration of data files. After the dataset is transformed into a clean dataset, the dataset is divided into training and testing sets so as to evaluate. Here, the training values are taken as the more recent values. Testing data is kept as

5-10 percent of the total dataset

3) Stage 3: Training Neural Network: In this stage, the data is fed to the neural network and trained the prediction is done by assigning random biases and weights. Our LSTM model is composed of a sequential input layer followed by 2 LSTM layers and 4 dense layers with ReLU activation and then finally a dense output layer.

4) Stage 5: Output Generation: In this layer, the output value generated by the output layer is compared to the target value from the dataset. The error or the difference between the target and the obtained output value is minimized by using back propagation algorithm which adjusts the weights and the biases of the network

1) Advantages of LSTM

Handling Long Sequences: LSTMs are well-suited for processing sequences of data with long-range dependencies. They can capture information from earlier time steps and remember it for a more extended period, making them effective for tasks like natural language processing (NLP) and time series analysis.

Avoiding Vanishing Gradient Problem: LSTMs address the vanishing gradient problem, which is a common issue in training deep networks, particularly RNNs. The architecture of LSTMs includes gating mechanisms (such as the forget gate) that allow them to control the flow of information and gradients through the network, preventing the gradients from becoming too small during training.

Handling Variable-Length Sequences: LSTMs can handle variable-length input sequences by dynamically adjusting their internal state. This is useful in many real-world applications where the length of the input data varies.

Memory Cell: LSTMs have a memory cell that can store and retrieve information over long sequences. This memory cell allows LSTMs to maintain important information while discarding irrelevant information, making them suitable for tasks that involve remembering past context.

Gradient Flow Control: LSTMs are equipped with mechanisms that allow them to control the flow of gradients during backpropagation. The forget gate, for example, can prevent gradients from vanishing when they need to be propagated back in time. This enables LSTMs to capture information from earlier time steps effectively.

2) Disadvantages of LSTM

Computational Complexity: LSTMs are computationally more intensive compared to other neural network architectures like feedforward networks or simple RNNs. Training LSTMs can be slower and may require more resources.

Overfitting: Like other deep learning models, LSTMs are susceptible to overfitting when there is insufficient training data. Regularization techniques like dropout can help mitigate this issue.

Hyperparameter Tuning: LSTMs have several hyperparameters to tune, such as the number of LSTM units, the learning rate, and the sequence length. Finding the right set of hyperparameters for a specific problem can be a challenging and time-consuming process.

Limited Interpretability: LSTMs are often considered as "black-box" models, making it challenging to interpret how they arrive at a particular decision. This can be a drawback in applications where interpretability is crucial.

Long Training Times: Training deep LSTM models on large datasets can be time-consuming and may require powerful hardware, such as GPUs or TPUs.

L. 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. SVM originated as an implementation of Vapnik‟s 1995 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 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.

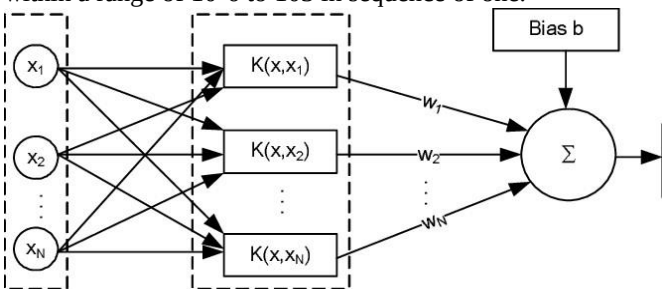

Figure 2.3 The architecture of the Support Vector Machine.

SVM Parameters

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

i. **Kernels** can be categorized as linear and polynomial kernels calculate the prediction line. In linear kernels prediction for a new input is calculated by the dot product between the input and the support vector.

ii. **Gamma parameter** measures the influence of a single training on the model. Low values signify far from the plausible margin and high values signify closeness from the plausible margin.

iii. C parameter is known as the regularization parameter; it determines whether the accuracy of model increases or decreases. The default value of c=10. Lower regularization value leads to misclassification.

M. Characteristics of support vector machine

i. SVM constructs a decision boundary with the largest possible distance to example

points (maximum margin separator) which helps it generalize well.

ii. It possesses the ability to embed data into a higher-dimensional space using kernel trick

iii. They combine the advantages of parametric and non-parametric models and have flexibility to represent complex functions.

Advantage, disadvantages of support vector machines

1) Advantages of SVM

One huge advantage of SVM is that it provides a globally optimized solution. SVM is effective in high dimensional spaces. It is effective in cases where number of dimensions is greater than the number of samples. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

2) Disadvantages of SVM

The main disadvantage of SVM is that it has several key parameters like C, kernel function and Gamma that all need to be set correctly to achieve best classification results for any given problem. The storage of variables requires a lot of memory space and computation time. It does not perform well, when we have large data set.

III. Evaluation Metrics

Several evaluation metrics can be used to assess the performance of an ensemble model combining Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) for stock price prediction. Commson metrics include:

A. Accuracy

The ratio of correctly predicted instances to the total number of instances provides a general measure of model correctness.

Accuracy is a measure of how close a measured value is to the true value or the actual value. It is usually expressed as a percentage. The formula for calculating accuracy is:

$$Accuracy = \frac{(Number\ of\ correct\ predictions)}{(Total\ number\ of\ predictions)}\ x\ 100\% \quad (2.1)$$

where the number of correct predictions is the number of times the model correctly predicted the outcome, and the total number of predictions is the total number of predictions made by the model (MerrianWebster., 2023).

B. Precision

The proportion of true positive predictions among all positive predictions, emphasizing the accuracy of positive predictions.

Precision is a measure of how many of the positive predictions made by a model are true. It is calculated as follows:

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})} \quad (2.2)$$

where True Positives are the number of correct positive predictions made by the model, and False Positives are the number of incorrect positive predictions made by the model.

C. Recall (Sensitivity)

The proportion of true positive predictions among all actual positive instances, focusing on the model's ability to capture positive instances.

D. F1 Scores

The harmonic mean of precision and recall, offering a balanced metric that considers both false positives and false negatives.

E. Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

A measure of the model's ability to discriminate between positive and negative instances across different probability thresholds.

F. Area Under the Precision-Recall Curve (AUC-PR)

Similar to AUC-ROC, this metric assesses the model's performance in terms of precision and recall.

G. Mean Absolute Error (MAE)

The average absolute difference between predicted and actual values provides a measure of prediction accuracy.

H. Mean Squared Error (MSE)

The average of the squared differences between predicted and actual values gives more weight to larger errors.

I. Root Mean Squared Error (RMSE)

The square root of MSE provides a measure of the typical magnitude of prediction errors.

J. Mean Absolute Percentage Error (MAPE)

The average percentage difference between predicted and actual values, offers a relative measure of accuracy.

K. Brier Score

A metric for assessing the accuracy of probabilistic predictions, particularly relevant for ensemble models providing probability estimates.

L. Confusion matrix

A table summarizing the number of true positive, true negative, false positive, and false negative predictions, facilitates a detailed analysis of model performance.

IV. METHODOLOGY

This chapter presents the details of Machine Learning (ML) algorithms adopted in this study and their implementation for predicting Nigerian Stock Exchange. Ensemble Long Short Time Memory (LSTM) method which combines decisions from multiple sub-models to a new model and then to make the final output to improve the prediction accuracy or the overall performance also with Support Vector Machine (SVM) model that utilizes input data and performs binary classification.

Predictive models

This study adopts three bases line ML algorithms, namely Ensemble learning, LSTM, and SVM, based on their superiority for ensemble learning in financial analysis.

A. System Architecture

Our experiment follows the basic data mining process demonstrated in the flowchart

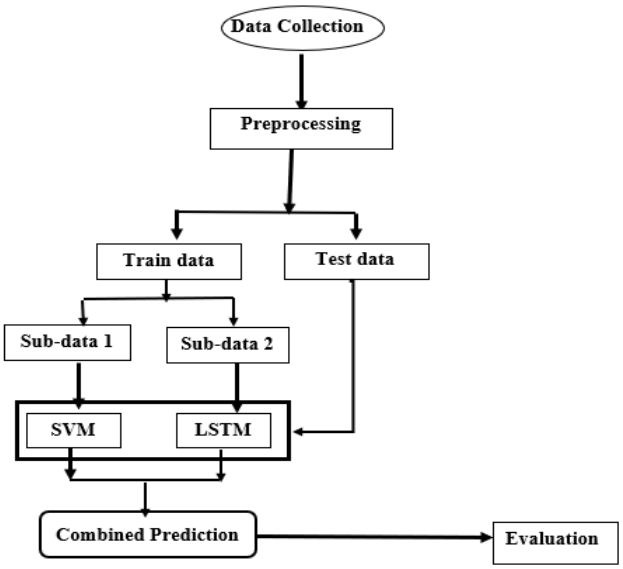

Figure 3.1 The Proposed Model Architecture

The overall flow of the proposed ensemble LSTM and SVM is demonstrated in Figure 3.1. The diagram as 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.

B. Data Collection

[13]. 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.

C. 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.

D. Data Preprocessing

Data preprocessing is a crucial step in preparing raw data for machine learning models. It involves cleaning, transforming, and organizing the data to make it suitable for analysis and model training. This step involves different activities in the analysis. Those actions like Replacing missing values, Dealing with outliers, Normalising, and scaling, handling categorical variables, Time series organization, Feature engineering, handling imbalance data, labeling for supervised learning, Data splitting etc, are sets of activities that can be carried out through the Data preprocessing stage and beyond.

Several techniques were employed for these actions respectively; This research considers Machine learning imputation for replacing missing data. In the case of an outlier, if found in the dataset, we employ the trimming values technique. Min-Max scaling is employed for Normalisation, to ensure features with larger magnitudes do not dominate the learning process. One hot encoding that creates binary columns for categorical variables. Lag Feature engineering will be used in dealing with time series data organization. The oversampling technique will handle imbalanced data where necessary.

E. Data Split: Training Set

The training set is a subset of a dataset used to train a machine learning model. It consists of input-output pairs, where the inputs are the features used by the model, and the outputs are the corresponding labels or target values. The model learns from these examples during the training process, adjusting its parameters to make predictions that minimize the difference between the predicted and actual outputs. In the context of stock market prediction or any other supervised learning task, the training set typically includes historical data. The machine learning model is trained on the training set by adjusting its internal parameters through an optimization process. The model aims to learn patterns and relationships within the training data that enable it to make accurate predictions on new, unseen data. Training involves feeding the input features (X_train) into the model, comparing the model's predictions with the actual labels (Y_train), and updating the model's parameters to minimize the prediction error.

F. Data Split: Testing Set:

Once the model is trained and validated, it is evaluated on a separate testing set that it has not seen before. This set helps assess the model's performance on completely new data and provides an estimate of its ability to generalize to real-world scenarios.

The training set is a critical component of supervised machine learning. It is used to teach the model the patterns and relationships present in historical data, enabling it to make predictions on new, unseen data. The quality and representativeness of the training set are essential factors in the success of the machine learning model.

V. ACKNOWLEDGEMENT

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

My heartfelt respect and gratefulness goes 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.

VI. 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.

REFERENCES

[1] R. Lekhani, Stock Prediction using Support Vector Regression and Neural Networks. International Journal of Advance Research, Ideas and Innovations in Technology, 2017. 3(6)

[2] O. Kowalewski and P. Śpiewanowski, Stock market response to potash mine disasters. Journal of commodity markets, 2020. 20, 100124.

[3] 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, 2020. 53(4), 3007-3057.

[4] I. Medarhri, M. N. Nouisser, F. Chakroun, K. Najib. Predicting Stock Market Price Movement using Machine Learning Techniques. In 2022 8th International Conference on Optimization and Applications (ICOA) (pp. 1-5). IEEE. 2022

[5] 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, 2021. 10(21), 2717.

[6] S. S. Liew, M. Khalil-Hani, and Bakhteri, "An optimized seR. cond order stochastic learning algorithm for neural network training," Neurocomputing, 2021. vol. 186, pp. 74-89.

[7] M.H. Kwayist, Nigerian Stock Exchange (NGX) Live. African x changes https://afx.kwayisi.org/ngx/ 2023.

[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 2019. (EICT) (pp. 1-6). IEEE.

[9] A. Durgapal, and V. Vimal, Prediction of stock price using statistical and ensemble learning models: a comparative study. In 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). 2021. (pp. 1-6). IEE

[10] Z. Fang, X. Ma, H. Pan, G. Yang, and, G. R. Arce, Movement forecasting of financial time series based on adaptive LSTM-BN network. Expert Systems with Applications, 2023. 213, pp. 119207.

[11] Y. Li, and Y. Pan, A novel ensemble deep learning model for stock prediction based on stock prices and news. International Journal of Data Science and Analytics, 202220, pp 1-11.

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

[13] 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