A Survey on Stock Price Prediction using Machine Learning Techniques

1st Abhinay Yadav
Department of Information Technology
Rajkiya Engineering College Ambedkar Nagar
Uttar Pradesh, India
abhinay80094@gmail.com

2nd Vineet Kumar
Department of Information Technology
Rajkiya Engineering College Ambedkar Nagar
Uttar Pradesh, India
vk28.rkt@gmail.com

3rd Satyendra Singh
Department of Information Technology
Rajkiya Engineering College Ambedkar Nagar
Uttar Pradesh, India
satyendrasinghjlp55@gmail.com

4th Ashish Kumar Mishra
Department of Information Technology
Rajkiya Engineering College Ambedkar Nagar
Uttar Pradesh, India
ashish.rcs51@gmail.com

2023 Winter Summit on Smart Computing and Networks (WiSSCoN) | 979-8-3503-3405-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/WISSCON56857.2023.10133865

***Abstract*—In Today's competitive environment, every Industry wants to grow fast to become a leader in their business. There is a need for the regular growth of business in the form of expansion. For Expansion, they required more capital. There are primarily 3 ways to raise Capital. These ways are an Initial Public Offering (IPO), Angel investors, and Business loans. Generally, when a company grows beyond a certain size, it is difficult for individual investors to continue to run the operations within their own capacity of capital. Companies need constant capital. An IPO method for raising capital often raises a company's profile and enhances its credibility with suppliers, Creators, and customers. Raising capital is the primary motivation for an IPO, however, it can also provide an opportunity for a partial sell-down by founders/early-stage investors. After IPO, the shares of a company are available in the form of stocks in an open platform. Including stocks in a public investment portfolio can prove to be beneficial. By investing in stocks of various companies, individuals can generate savings and shield their money against inflation and taxes. However, in order to optimize returns on investment, it is important to have the ability to predict stock prices. All the surveyed techniques are based on the concepts of Machine Learning. These techniques are compared by highlighting their strengths and weaknesses. In the survey of existing works, it is found that combining LSTM with another model can be the most effective technique for stock price prediction. The future scope of improvements in the surveyed research is also suggested in the paper. Researchers in this field can take advantage of these suggested improvements for further research in the field to enhance the performance of techniques used for stock price prediction.**

***Keywords*—Stock Market, Trading, Investment, Machine Learning(ML) Algorithm, Prediction, LSTM, SVM, GRU, ARIMA, ARMA, etc.**

I. Introduction

The most critical challenge in today's financial market is to find an efficient method for the prediction of stock prices. These methods can be utilized for maximizing profit from the investment. And the techniques can be used to forecast future prices based on existing information and stock sentiments. For many individuals, forecasting the Stock price is a challenging task. Researchers from several fields, including computational business and economics, are conducting research on stock market forecasts. They have explored a range of approaches for forecasting the market, including diverse tactics and algorithms, as well as a combination of indicators like company-related news, Investor sentiment, Politics, and Exchange rates factors. In the following subsections A, B C, the description of the Share Market, Market Analysis, and Investment Vs Trading, have been provided respectively.

A. Information About Share market

One can purchase and sell shares of publicly traded corporations on a stock exchange, which is a marketplace that is open to the public. Equities, which are also referred to as stocks, represent ownership interests in a corporation. The stock exchange acts as a middleman, facilitating the purchase and sale of stocks. The stocks are associated with public sector companies and the firm. stocks can be also called shares that are commonly used in everyday speech. People refer to it as an investment plan and hold stocks for a long time in a delivery position that secures the abundance of cash throughout retirement. The market is unpredictable, but by using algorithm trading one can predict stocks up to some extent, Some approximation and prediction methods have been developed namely approximate values and rough numbers are created in the hope of the best results, prediction can be done using the dataset like historical data, sentiment data, and news headlines.

B. Stock Market Analysis

Share market forecasting and assessment are two of the most challenging tasks. Market values of the stocks are influenced by a variety of factors. These factors can be dependent factors (i.e. average value, Stock news) or independent factors (i.e. market sentiments). Due to these factors, stock prices are fluctuations [9]. These factors make it very difficult for the trader or investor to identify the stock's trend (downward or upward) with great accuracy. There are generally two kinds of forecasts used in stock market prediction systems. These are dummy forecasts and real-time forecasts. Dummy forecasts created some Dummy forecasting concepts and projected the upcoming price of the shares on the basis of the average price. To produce a real-time forecast, one must connect to the internet and examine the present share price of the firm.

C. Investment vs. Trading

According to Warren Buffett, "if you don't find out how to make money while you sleep, you'll work till you die". Investing occurs when you acquire stock with a long-term view. As the firm expands the value of the investment increases. This is referred to as passive income. The investment continues to grow even when you are on vacation.

TABLE I
INVESTMENT VS TRADING

Investment	Trading
Short-term	Long-term
Profit mindset	Growth mindset
More risk involved	less risk involved
Benefits from volatility	Benefits from stability
More technical analysis	More fundamental analysis
Best for building capitals	Best for growing capitals
Buying/selling market waves	buying and holding

When compared to trading, investing is a very simple game. Trading needs advanced market knowledge, real-time research, and the ability to identify price changes in fractions of a second. Table 1 shows the difference between investment and trading. Retail investors who wish to get passive income without dedicating a lot of time to research should invest their money for a long duration. As an investor, one has a better chance of increasing capital. Trading can be attempted by someone who has adequate information and a keen sense of the market. Figure 1 shows a technique, that must be followed during trading.

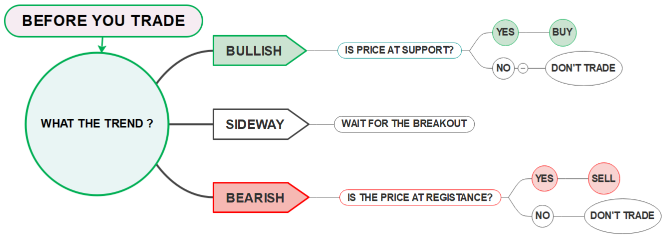

Fig. 1. How to Trade

D. Price-to-Earnings (P/E) ratio

The price-to-earnings (P/E) ratio is a valuation indicator used to compare the value of a company's shares to its earnings. It is computed by dividing a company's current market price by its profits per share (EPS). In other words, it is a measure of how much investors are willing to pay for each rupees of a company's earnings.

The P/E ratio is often used by investors to assess whether a stock is overvalued or undervalued relative to its earnings potential. A high P/E ratio suggests that shareholders have great expectations for the business's future profit growth, whereas a low P/E ratio indicates that the company is possibly undervalued and set for growth. However, it is important to note that P/E ratios can vary widely across industries, so it is generally more useful to compare a company's P/E ratio to its peers in the same industry. Additionally, the P/E ratio should be used in conjunction with other valuation metrics and fundamental analysis to make informed investment decisions.

The rest of the article is organized as follows: Section II presents the literature Survey and a Detail of machine learning techniques used for the prediction is provided, Comparison is done in section III, State-of-the-art research directions are illustrated in Section IV, Finally, Section V concludes the paper.

II. LITERATURE SURVEY

For the majority of us, "what other people think" has always been a crucial piece of knowledge when making decisions [9]. The World Wide Web nowadays permits us to find out about the viewpoints and experiences of the public who are neither personal associates nor well-known skilled reviewers – in different words, people have got never heard of them. As a result, a rising variety of individuals are expressing their views with strangers over the Internet. Individual customers' interest in online opinions on products and services, similarly because of the potential impact of such views, maybe a driving part of this field of research. during this procedure, there are various barriers to overcome.

There are several further problems with this procedure that must be addressed so as to induce the required outcomes. There has been a rise in interest in employing automated trading systems to anticipate stock prices and create winning transactions in recent years. Mariam Moukalled, Wassim El-Hajj, and Mohamad Jaberhave suggested a revolutionary trading strategy that combines mathematical functions, advanced analytics, and external elements such as news emotions to enhance stock price forecast accuracy and boost trading decisions. Utilizing data from the first few market hours of the day, the method attempts to forecast the price and trend of a certain stock by the close of the day. In their approach,The researchers employed both traditional machine learning approaches and a variety of deep learning models that considered the relevance of important news. They conducted various experiments to compare the effectiveness of different models in predicting the direction of today's close price relative to yesterday's close price based on the features.

The results showed that SVM achieved the highest accuracy of 82.91% for Apple Inc. (AAPL) stock. Aparna Nayak,

M. M. Manohara Pai, & Radhika M. Pai have suggested a novel method for forecasting market movements that utilizes both historical prices and sentiments extracted from social media data and news. Their models, created using supervised machine learning algorithms, demonstrate an impressive accuracy of up to 70% for both daily and monthly predictions.

On the other hand, T. Manojlović and I. Štajduhar suggest a distinct method of Just used the random forests method to create forecast models for stock market activity five and ten days ahead. Their models incorporate technical indicators and achieve an average classification accuracy of 76.5% and 80.8% for 5 and 10-forward models, respectively, as evaluated using 10-fold cross-validation. In this regard, C.H. Vanipriya1 and K. Thammi Reddy have suggested a novel machine learning-based share price reliable indicator using neural networks that combines both historical pricing models and sentimental analysis to enhance accuracy.

The proposed system inputs historical prices and sentimental values into a hybrid neural network to increase the efficacy of share price estimation techniques. Lexicon approaches and techniques for machine learning are the two basic ways of obtaining emotions. The researchers have employed a supervised training set with a maximum of 1000 iterations, a learning rate of 0.7, and a maximum error of 0.0001 to train the model. Training ceases after reaching a total net error below 0.0001.

Similarly, Edgar P. Torres P, Edgar A. Torres Hernández, and his colleagues have suggested a method for forecasting stock market values that makes use of actual information and machine-learning techniques. To estimate closing prices, historical data of the highest price, the researchers used random trees and Multilayer perceptron algorithms. stocks obtained from Google Finance. The variable close has been chosen for prediction since it represents the latest price at which all participants agreed for economic freedom in the traded session. This strategy, however, is primarily reliant on the attitudes, sensations, and intentions of respondents. The researchers have employed the WEKA packages to execute the proposed model. Pritam Ahire, Hanikumar Lad, Smit Parekh, Saurabh Kabrawala, and D.Y Patil, all are involved in the reachers that focus on the use of a hybrid model consisting of LSTM with RNNs (hybrid) machine learning technique that anticipates stock values more accurately. Stock costs for corporations adore metallic element Corporation, Carnival private company, Tesla private company, et al. are forecasted by the algorithm. for every firm, the approach relies on 5 factors: the stock value date, the beginning price, the lowest price, the volume, and the shutting price.

In the work done by Ashish Pathak, and Nisha P Shetty they recommend that the model be improved by introducing finer fuzzy rules. rising the dimensions and length of training data may end up in higher prediction. Their model is based on Machine learning and sentiment analysis, A forecasting model has also been developed using the given approach to estimate real returns or investments in a timely manner. this might be wont to illustrate the model's accuracy. This formula will propose the most effective stocks for investment. The study done by V Kranthi Sai Reddy, proposes that their analysis explains the way to predict a stock using Machine Learning. Most stockbrokers create stock forecasts using technical and fundamental analysis or statistical research. The programing language Python is used to forecast the monetary markets using machine learning. The users have a tendency to propose an (ML) approach that will be taught using publicly accessible stock information to achieve intelligence and so use that knowledge to come up with a correct forecast during this study.

The study was done in Korea by Hyeong Kyu Choi, In this study, they use the Support Vector Machine (SVM) technique and take small as well as large company data during the training phase, in order to optimize a portfolio, They observed that it's essential to forecast the value correlation of 2 assets across future time frames. They anticipate the stock price parametric statistics of two distinct stocks' using LSTM and RNN algorithm combine this is also a type of hybrid model consist both LSTM and RNN models. RNNs have the ability to know temporal dependencies. Any enhancing its long-term prediction abilities are the LSTM cells. In order to take into consideration the model's linear and non-linearity, they additionally apply the ARIMA model. The LSTM model receives the residual value from the ARIMA model, Which filters knowledge for linear trends.

On the other hand, the study of research done by Hongming Wang, College of Information and Control Engineering, Qingdao University of Technology is also done, He conducted research to develop a stock selection technique in machine learning based on the Logistic Regression model and the SVM model. Both concepts have had widespread applications in a variety of sectors. The Logistic Regression and SVM models are both capable of accurately forecasting the stock market. Both could be used to choose a decent enough investment portfolio to achieve an objective rate of return. The SVM model outperformed the Logistic Regression model in terms of return and maximum Drawdown in the model we created. Furthermore, investment strategies employing the Ridge Regression and SVM models exhibited a greater excess return rate. It was also superior to the performance of the stock market index at any moment.

The research published in IEEE 2019 by Jeevan B et. al. discusses the topic of Market Prediction Using ML Technique, In this work, he concludes that the stock market has been the talk of the town, with an increasing number of academics and business people displaying interest in it. This study focuses on the technique of forecasting stock prices on the National Stock Exchange using Hybrid model that consists of both RNN and LSTM, using a variety of data such as the current market price and anonymous events. This study also mentions a recommendation system, as well as models built on RNN and LSTM approaches, that are employed in picking the firm.

In the hybrid model used by authors namely Ya Gao, Rong Wang, and Enmin Zou at, University in Xi'an, China, research is based on the hybrid model i.e. They have demonstrated that both techniques, LSTM and GRU, can forecast stock values efficiently. In this reachers, they apply various indicators like technical indicators, including investor mood indicators.They analyze market sentiment also during model training.The effectiveness of LSTM+GRU for prediction of the stock exchange under numerous conditions are compared in this paper. The paper was published by Neha Bhardwaj, MD Akil Ansari students of VIT University, They work on multiple different models and after comparison, they conclude that logistic regression offers the ability to forecast and analyze market movement direction more precisely. A variety of models, including Autoregressive Integrated Moving Average and Random Foresthave also become well-known in stock market forecasting. Random Forest has proven to be useful in categorization work, ARIMA on time series prediction, and financial applications. In the experiment, the K-NN model is also used, and it produces some promising results in forecasting stock market trends.

In machine learning, the "No Free Lunch" theory indicates that there is no one optimum approach for all situations, particularly in supervised learning. Consider parameters like as dataset size and structure, then experiment with appropriate strategies on a test set. The research focuses on identifying critical approaches and stages for achieving good outcomes in machine learning while avoiding plagiarism and appropriately attributing sources.

A. Support Vector Machines

It's an application of perceptron-based machine learning. Since they are organized in a vector space, Perceptron cannot independently learn their patterns. Support vector machines expand vector space by allowing patterns from lower-level to higher-level vector spaces. The support vectors—new vectors that are associated with certain subpatterns are used to do the above. A support vector machine's objective is to discover the support vectors required to understand a pattern.

In stock price prediction, the Support Vector Machine (SVM) method is often utilized. It is a supervised learning method that makes predictions using labeled data. SVM determines the optimum hyperplane to divide data points into classes. Using mathematical principles such as kernel functions, SVM predicts whether the stock price will climb or decline. Kernel functions are used to find patterns and correlations by transforming data into higher-dimensional feature spaces. The proper kernel function must be used for reliable SVM predictions..

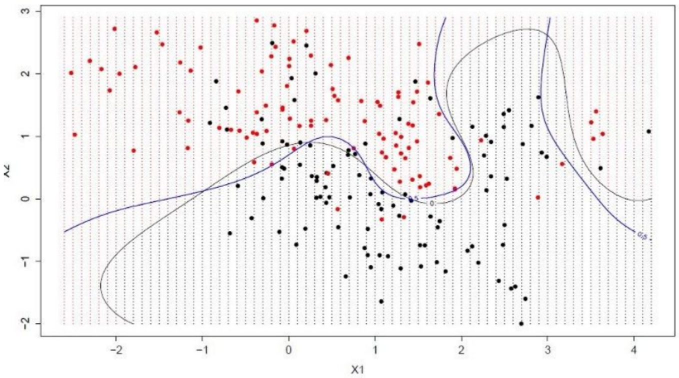

Fig. 2. SVM

There are several types of kernel functions that can be used with SVM for stock price prediction, For good predictions, the SVM algorithm transforms input data into higher dimensional feature spaces using various kernel functions. In the original feature space, the linear kernel locates the hyperplane. Polynomial functions are used by the polynomial kernel to translate data into a higher dimensional feature space. Using a Gaussian function, the RBF kernel turns data into an infinite-dimensional feature space. Using a cost function to punish misclassifications and determine the best feasible hyperplane, the SVM algorithm then identifies the ideal hyperplane that divides distinct classes of data points.

The mathematical formula for the SVM cost function is:
minimize $1/2||w||^2 + C_i$
subject to $y_i(w^T x_i + b)1-_{i,i} 0$

For stock price prediction, the SVM (Support Vector Machine) technique employs a weight vector (w), input data (x), a target variable (y), a bias term (b), a slack variable ($_i$), a regularisation parameter (C), and the squared norm of the weight vector ($||w||$2). It is critical to select the right kernel function and hyperparameters for accurate predictions. SVM may not be appropriate for huge datasets and may need substantial processing resources.

The Support Vector Machine (SVM) is used during the implementation of classification and regression techniques. SVM is based on a supervised machine learning algorithm. However, classification problems are those that are most frequently applied. This approach represents each data point as a point in an n-dimensional space, where n is the total number of attributes. Each feature's output is associated with a specific coordinate value. Then the classification is done by determining the coordinates of a single observation that serves as the hyperplane. This hyperplane. This hyperplane effectively separates the two classes. Fig. 2 shows Hyperplane dividing the n-dimensional space support Vectors [11]. This method best separates the two classes (hyper-plane/line).

B. Regression Techniques

Regression is a machine learning technique that may be trained to forecast outcomes with real numbers, such as temperature, stock price, etc. A hypothesis that may be linear, quadratic, polynomial, non-linear, etc. serves as the foundation for regression. The hypothesis is a function that depends on some input values and hidden factors. The hidden parameters are adjusted in the training phase based on the input values given. The gradient descent algorithm is the mechanism that performs optimization. In order to compute the gradient at each layer when using neural networks, the back-propagation method is also required. The same hypothesis is used to predict outcomes that will once more represent real values once the hypothesis factors have been trained, using new input data and the learned factors (when they delivered the least error during the training).

C. LSTM

Hochreiter and Schmidhuber invented the LSTM (Long Short-Term Memory) recurrent neural network (RNN) in 1997. It solves the vanishing gradient problem that typical RNNs have while processing large data sequences. LSTM cells have gates that govern the flow of information as well as a memory cell that can retain data for extended periods of time. The input gate determines whether fresh information is added to the memory cell, the forget gate determines if the information stored in the memory cell is maintained or deleted, and the output gate determines how much information is output.

The LSTM formula for one time step can be expressed as follows:

Input Gate

i(t) = sigmoid(W$_i * [h(t-1), x(t)] + b_i)$

Forget Gate

f(t) = sigmoid(W$_f * [h(t-1), x(t)] + b_f)$

Output Gate

$o(t) = sigmoid(W_o * [h(t-1), x(t)] + b_o)$

Memory Cell

$c(t) = f(t) * c(t-1) + i(t) * tanh(W_c * [h(t-1), x(t)] + b_c)$

Hidden State

$h(t) = o(t) * tanh(c(t))$

where:

The equation describes a neural network with recurrent operation that takes input $x(t)$ and hidden state $h(t-1)$ from the previous time step and uses three gate functions $i(t)$, $f(t)$, and $o(t)$ to determine how much relevant data to maintain or ignore from the input signal and hidden layer state, and the amount to outcome to the existing hidden layer. Input, forget, and output gates, as well as prior memory cell state, weight matrices (Wi, Wf, Wo, and Wc), and bias vectors (bi, bf, bo, and bc) [18]. are used to determine current memory cell state $c(t)$. Tanh and sigmoid are activation functions. For many time steps, the LSTM equation requires supplying the preceding step output as input to a current time step.

The primary advantage of LSTM is its ability to retain intermediate information over a prolonged period. Unlike traditional recurrent neural networks, LSTM units have memory cells that store information over a long or short duration without relying on the activation function applied in the repeating components [4]. An crucial point is that any cell state may be replicated only by releasing the forget gate, which has a value between zero and one. That is, The forgetting approach within the LSTM cell is up to speed with each hardware and therefore the operation of cell state activation. Rather than explicitly growing or reducing in each step or layer, the previous cell's knowledge may accept the unrestrained cell, and the instrument will change to its applicable levels within a limited time. Because the amount decrease within the memory block does not renew in an extremely revenant manner, the gradient does not end once instructed to spread backward.

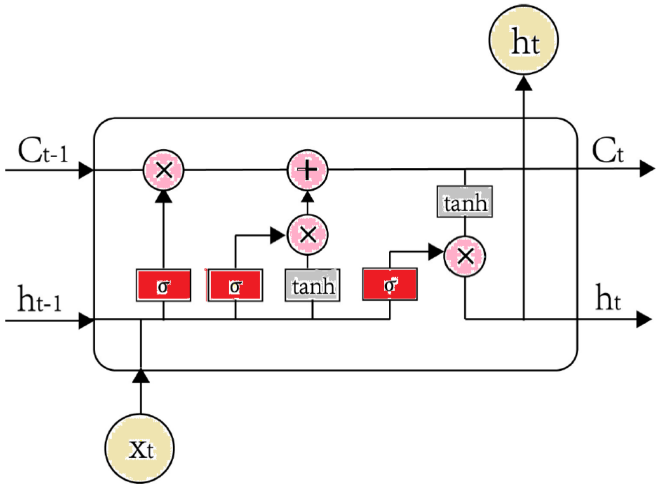

Fig. 3. LSTM

D. Gated Recurrent Unit (GRU)

The GRU is an of type of RNN that is largely beneficial over lengthy short-term memory in certain types of situations (LSTM) [7]. It consumes less memory and is quicker than LSTM; yet, The accuracy of LSTM is high when dealing with larger sequence datasets. single cell GRU structure is provided in Fig. 4 [13].

GRUs also handle the vanishing gradient problem (values used to update network weights), which is a difficulty with traditional recurrent neural networks. Grading may become too little to affect learning if it shrinks over time as it back propagates, rendering the neural net untrainable. RNNs can effectively "forget" lengthier sequences if a neural net layer is unable to learn.GRUs address this issue by utilizing two gates: an update gate and a reset gate. These gates control what information is given to the output and may be trained to hold data for an extended period of time. this permits it to convey essential information in a few succession of occurrences so as to produce further correct forecasts.

1. Update gate: The update gate in a Gated Recurrent Unit (GRU) has a similar function to an LSTM's forget and input gates, as it determines which information to discard and which to keep for the current time step. This gate is crucial in deciding how much information from previous time

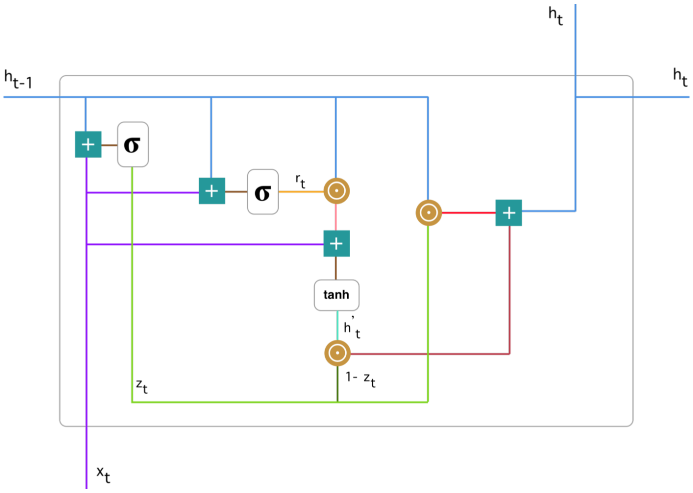

Fig. 4. Gated Recurrent Unit

steps should be propagated to the future. Unlike some other models that may encounter the vanishing gradient problem due to losing important historical information, the GRU's design ensures that all past data is retained, making it a powerful solution for processing sequential data.
2. Reset gate: A reset gate is a form of gate utilized by the model to determine how often past data to discard.

The use of GRUs and LSTMs in deep learning models is a common practice, and there is ongoing debate about which one is better. While GRUs have fewer tensor operations, allowing for faster training, there is no clear winner in terms of overall performance. Researchers and engineers usually experiment with both types to determine which one is the best fit for their specific application.

One advantage of GRUs is that they do not require a memory unit to regulate the flow of information like LSTMs do. They can directly use all hidden states without any control. This means that GRUs may require fewer parameters and, therefore, less training data to generalize accurately.

However, when presented with vast amounts of data, LSTMs may produce superior outcomes due to their increased expressiveness. Ultimately, the choice between GRUs and LSTMs depends on the specific use case and the available data. Researchers and engineers must carefully evaluate both options to determine which one is the best fit for their needs.

The formulas for a single timestep of a GRU model are:
Reset gate:
$r_t = (W_r * x_t + U_r * h_{t-1} + b_r)$
Update gate:
$z_t = (W_z * x_t + U_z * h_{t-1} + b_z)$
Candidate hidden state:
$h_t^{'} = tanh(W_h * x_t + r_t * (U_h * h_{t-1}) + b_h)$
Hidden state:
$h_t = (1 - z_t) * h_{t-1} + z_t * h_t^{'}$
where:
In an LSTM-based RNN, input is xt at time t, hidden state is $ht - 1$ at time t-1, reset gate is rt at time t, and update gate is zt at time t. At time t, the contender hidden state is ht', and the final hidden state is ht. Weight matrices and bias vectors (W, U, and b) are learnt during development. $sigma$ is the sigmoid activation function, while tanh is the hyperbolic tangent activation function.

Using the input vector, previous unit output, and sigmoid activation function, the GRU unit computes the update and reset gates. In the candidate value computation, the reset gate is utilized to determine how much information from the prior state should be maintained. The updating gate, which calibrates the prior output and the current candidate output, is used to compute the output value. Both LSTM and GRU require an intermediary gating mechanism to calculate output values, and their performance varies depending on the job [17]. GRU beats LSTM in several tasks, although LSTM outperforms GRU in voice recognition [18]. Yet, because GRU has fewer parameters than LSTM, it can optimise quicker. Both models are strong and well-suited to sequential data in general [19].

E. Moving Average (MA)

Average is the foremost ordinarily used words in our daily life. computing the typical marks to assess whole performance, and determining the average prices of the last days to induce a sign of today's value - these are all common jobs that we tend to conduct on a daily basis [4]. As a result, this is often a solid place to begin generating predictions on our dataset. every day's forecasted terms are going to be the average of antecedently determined values. rather than a straightforward average, we are going to use the MA approach, that uses most recent set of variables for every forecast. In different words, for every consecutive step, the projected prices are thought of whereas the oldest determined value is off from the collection. Fig. 5 shows the Stock price moving average trend [14].

Moving average is a commonly used method for stock price prediction that can help smooth out short-term fluctuations and identify longer-term trends. The moving average is calculated by taking the average price over a certain number of time periods and then shifting the window forward one period at a time.

The formula for a simple moving average (SMA) for a given time period n is:

$$SMA_n = (P_1 + P_2 + ... + P_n)/n$$

where:
SMA_n is the simple moving average for time period n
$P_1, P_2, ..., P_n$ are the stock prices for the last n time periods
For example, if a person wanted to calculate the 10-day moving average for a stock, it would take the average of the last 10 days of prices. The next day, one would drop the oldest price and add the newest price, and recalculate the moving average based on the new set of 10 prices.

An exponentially moving average (EMA) has a similar approach, but it gives more importance to recent values than to previous prices. The formula for computing an EMA for a given time period n is as follows:

$$EMA_n = (P_n * (2/(n+1))) + (EMA_{n-1} * (1 - (2/(n+1))))$$

where:

EMA_n is the exponential moving average for time period of n.

P_n is the stock price for the current time period

EMA_{n-1} is the exponential moving average for the previous time period

The smoothing factor $2/(n+1)$ determines the weight given to the current price relative to the previous EMA. The smaller the value of n, the more weight is given to recent prices, and the more sensitive the moving average is to short-term fluctuations. Conversely, a larger value of n will give more weight to older prices and provide a smoother average that is less sensitive to short-term fluctuations.

Moving averages are often used in combination with other technical indicators, such as the Relative Strength Index (RSI) or Moving Average Convergence Divergence (MACD), to make trading decisions.

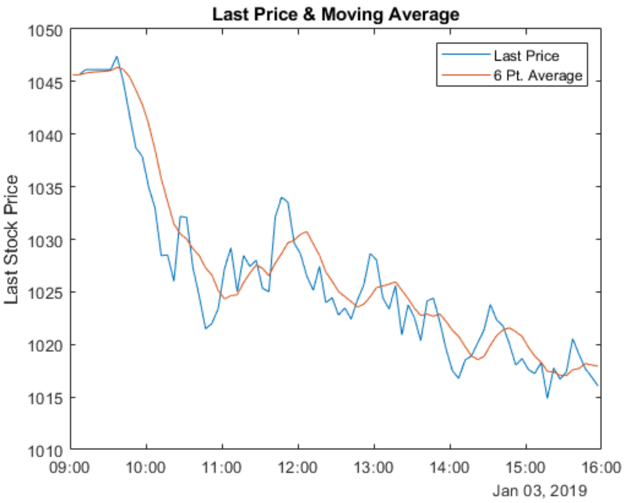

Fig. 5. Moving Average

F. Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a type of neural network architecture that are particularly well-suited for sequential data, such as time series data like stock prices. RNNs use internal memory to process input sequences of varying lengths and can capture complex temporal dependencies in the data.

Traditional neural networks, on the other hand, believe that inputs and outcomes are unrelated; The recurrent network's output is determined by the preceding attributes in the chain. The Hidden state is a state that can remember certain sequences of data, the core and most essential aspect of RNN.

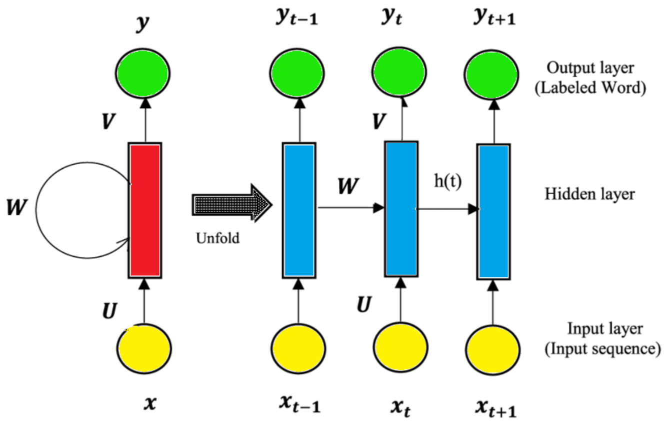

Fig. 6. Recurrent Neural Network

The basic equations for a single time step of a simple RNN are:

Hidden state calculation:

$$h_t = tanh(W_h * x_t + U_h * h_{t-1} + b_h)$$

Output calculation:

$$y_t = softmax(W_y * h_t + b_y)$$

where:

xt represents the input at time t, $ht-1$ represents the hidden state at time t-1, and ht represents the current hidden state at time t. The weight matrices Wh and Uh represent the input and hidden states, respectively. The bias vector for the hidden state is bh, and the hyperbolic tangent activation function is tanh. Wy is the output's weight matrix, and by is the output's bias vector. Softmax is a function that transforms the output to a probability distribution across all possible values.To train an RNN for stock price prediction, One would feed historical stock price data into the network as a sequence of inputs and train it to predict the next day's stock price. The output of the RNN is a probability distribution over possible future prices, so one would typically take the mean or median of the distribution as a predicted price.

In practice, more advanced variants of RNNs, such as LSTMs and GRUs, are often used for stock price prediction due to their ability to capture longer-term dependencies and avoid the vanishing gradient problem. The equations for LSTMs and GRUs are more complex than the simple RNN equations above, but they follow a similar structure of updating the hidden state based on the input and previous hidden state.

G. K-Nearest Neighbours

KNN is another intriguing ML method that may be used here (k nearest neighbors) [8]. KNN compares the similarity of new and old data points based on independent factors, Fig. 7. shows K-Nearest Neighbours [16]. KNN is among the most basic and widely used categorization strategies in machine learning. It is based on the supervised machine learning technique algorithm and is used in many fields like data mining, face stocks an other pattern recognition and intrusion detection. Because of its quasi-character, which implies that it does not make any fundamental assumptions about the distribution of data, it is commonly disposable in practical situations (unlike other algorithms, like GMM, which supposed a Gaussian distribution of the input data).

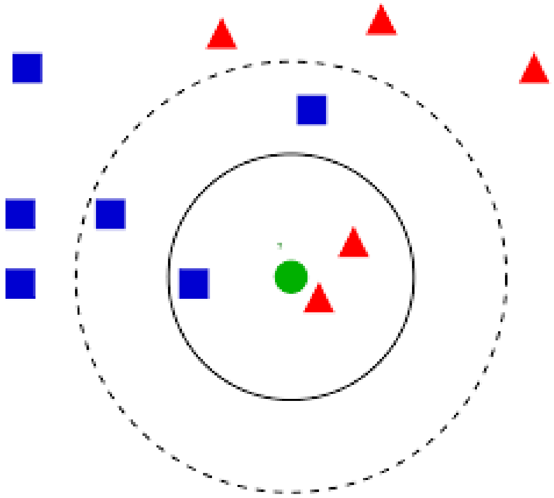

Fig. 7. K-Nearest Neighbours

H. ARMA

In time series, we frequently rely on historical data to anticipate present and future values. However, this is not always sufficient. Unexpected occurrences such as natural catastrophes, financial crises, or even wars can cause an abrupt shift in values. That is why we want models that can use prior data as a foundation for forecasts while also swiftly adjusting to unexpected shocks. Assume that "Y" is a random time-series variable. A basic Autoregressive Moving Average model would thus look like this:
The name ARMA stands for Autoregressive Moving Average. It is the result of combining two simpler models: the Autoregressive (AR) and the Moving Average (MA). Because we usually insert the residuals at the end of the model equation in analysis, the "MA" section comes second.

$$y_t = c + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t$$

AR model uses historical data to forecast future values. yt and yt-1 represent current and previous period data respectively, while $\epsilon t - 1$ and ϵt are the error terms. The error term from the previous period helps in improving the accuracy of the forecast. "c" is a baseline constant factor.

$\phi 1$ and $\theta 1$ are the parameters that represent the influence of the previous period's value and the prior error period, respectively, in understanding the current period's value in the ARMA model. In advanced models, ϕi and θi show the importance of values and errors for the "i-th" lag. $\phi 4$ indicates how much of the value from four periods ago still matters today, while $\theta 3$ represents the significance of the error from three periods ago for the current period.

I. ARIMA Model

An ARIMA model stands for AutoRegressive Moving Average (ARIMA). It is often used in statistical analysis to forecast future values or to higher perceive current information. In things, once data exhibits some non-stationarity, ARIMA models are used. Trends, cycles, random walks, and combos of the 3 are samples of non-stationary behavior. information points that don't seem to be stationary are surprising and can't be expected or modeled. Analysis performed on data from non-stationary time series might not be acceptable since it should counsel an affiliation between a pair of variables when only one of them is present. The non-stationary data should be born-again into stationary information so as to get repeatable, dependable findings. In contrast to the non-stationary approach, the stationary process reverts to a predefined long-term mean and exhibits constant variation over time, which combines a dynamic variance and a means that doesn't consistently approach or return to a long-term mean. The model's goal is to forecast future changes in the value of securities or the situation of financial markets by highlighting disparities between values within the series rather than the actual values themselves.
Autoregression (Ar) a model that accounts for the connection between an observation and a specific quantity of lag observations.
Integrated(I) The distinction between the information values and also the previous values is diagrammatical by Integrated(I), It shows that the raw data must differ for the statistic to become stable.
Moving Average(MA) is a type of model that considers the dependence among observation and a fixed range of delay observations.

III. Comparison of Prediction Techniques

A summarized comparative study of different algorithms is presented in Table II. The advantages and disadvantages of several algorithms are also provided for researchers in the field. The existing research work used several machine learning algorithms like LSTM, GRU, RNNs ARIMA, SVM etc. The list of used algorithms along with their corresponding pros and cons is highlighted in the table.

TABLE-II
COMPARATIVE STUDY OF DIFFERENT ALGORITHMS

TITLE	TECHNIQUE USED	PROS	CONS
A New Model for Stock Price Movements Prediction Using Deep Neural Network [34].	LSTM and GRU	Used hybrid model LSTM+GRU	Does not uses sentiment data
Stock Market Prediction Using Machine Learning [20].	• SVM	• SVM is more effective in high-dimensional spaces. • memory efficient	SVM does not function well when the set of data contains more noise, i.e. target categories overlap.
Empirical evaluation of gated recurrent neural networks on sequence modelling [17].	LSTM and GRU	•For both Ubisoft datasets, (GRU-RNN and LSTM-RNN) clearly beat the more typical tanh-RNN. • Comparison is done b/w GRU-RNN and LSTM-RNN	• Address the more difficult challenge of raw speech signal modeling, and refrain from reaching a firm judgment on which of the two gating units was superior.
A Stock price prediction method based on deep Learning Technology [35].	The Doc2Vec with LSTM models.	Combination of two datasets Historical + sentiment data.	• The social media text data obtained is from a single platform, and just one stock is chosen for prediction in this study.
An Empirical Exploration of Recurrent Network Architectures [18].	RNNs, LSTM and GRU	GRU is better than LSTM except for language modeling. LSTM with a large forget bias performed better than both LSTM and GRU on most tasks.	•Failed to find significantly different architectures. The top-performing architectures were similar to GRU. •Longer search is costly but necessary to find diverse architectures. Failure to improve significantly over LSTM suggests that better architectures are not easy to find.
LSTM-Based Stock Price Prediction [1].	LSTM	The main advantage of LSTM is its ability to read the intermediate context.	Does not uses news analysis data.

TABLE-II (Contd…)

TITLE	TECHNIQUE USED	PROS	CONS

Stock Price Prediction using Bi-LSTM and GRU based Hybrid Deep Learning Approach [32].	Bi-LSTM, LSTM and GRU,NN	• Optimal parameters for the proposed Bi-LSTM-GRU model: window size of 60 and a train-test splitting ratio of 65.
		• The model achieved the lowest errors:
		MSE =0.0000180
		RMSE=0.0042403
		MAE=0.0028890
		MSLE=0.0000118, and MAPE 28.058, when compared to other state-of-the-art models (Bi-LSTM, LSTM, GRU, and NN) using the same dataset.
			Does not considered both low frequency and high dimensional data while developing the model.

Predicting Stock Market Trends Using Random Forests [16].	Random Forests	• Suitable for Intraday Trading.	Give the wrong results when stock is influenced by politics or Conflict situations.
Indian Stock Market Prediction Using Machine Learning and Sentiment Analysis [2].	• Machine Learning Module • Sentiment Analysis module • Fuzzy logic Module	Combination of two datasets Historical +sentiment data.	• Stock news headlines are not covered • Only one stock is selected for prediction in this study
Automated Stock Price Prediction Using Machine Learning [25].	SVM and other external factors such as news sentiments	• The AI framework uses DNN, RNN, SVR, and SVM for prediction. • Test is done on APPL, AMZN, GOOGL, and FB stock shares and achieved 82.91% accuracy.	Does not try different time frames for grouping data to enhance prediction to exact price.
Stock Market Prediction Using Machine Learning [33].	Regression and LSTM	The LSTM model has shown good efficiency, resulting in positive outcomes.	Larger datasets can further improve the accuracy of stock market prediction systems.
A Combined Model of ARIMA-GRU to Forecast Stock Price [36].	ARIMA, SVM, and GRU	The hybrid ARIMA-GRU model delivers improved accuracy on a big dataset as a comparison to machine learning. Stock-SBIN Accuracy-SVM:72.12% GRU:94.60%	SVM alone does not perform well in large datasets.

IV. State of art Research Directions

The stock market prediction seems like a complicated problem because there are various factors that are still left unaddressed and do not seem to be Organized at first. In the above survey majority of authors have used only historical data. In addition to past prices, other relevant information such as politics, economic progress, financial news, the P/E ratio of stocks, and public sentiment may also have an impact on stock prices. Since it has been demonstrated in this digital age that sentiment (mood) analysis and stock news have a significant influence on the Stock pattern (bear/bull). As a result, combining technical and fundamental analysis can yield a very efficient prediction.

The existing models assessed used diverse machine learning or deep learning methodologies [7]. Predicting the price-to-earnings (P/E) ratio of a stock is a critical aspect of stock price prediction. The P/E ratio is a widely used valuation metric that enables investors to assess whether a stock is overvalued or undervalued compared to its earnings. Thus, it is essential to select an appropriate setup for P/E ratio prediction, taking into account the implications of various factors on prediction performance.

V. Conclusion

The purpose of this survey is to study different traditional methods, machine learning algorithms (ML), and deep learning algorithms that are used in stock market prediction. In the past, many scholars have used the above technique in predicting the stock market and achieved good results. From the literature survey done, It is found that one of the best algorithms for forecasting the market price is the hybrid machine learning model i.e. combination of LSTM and GRU, and for a more accurate result, The combination of three datasets. historical data, social media market sentiment data, and news headlines or news articles can be used. Along with this, the price-to-earnings (P/E) ratio is used to evaluate a company's stock price relative to its earnings. It is a useful tool in stock valuation, it should be used in conjunction with other factors. Hence, it can be concluded that parameters, investors, and traders will greatly be benefited from the survey to invest in the stock market.

References

[1] Pritam Ahire, Hanikumar Lad, Smit Parekh, Saurabh Kabrawala, "LSTM Based Stock Prediction," International Journal of Creative Research Thoughts(IJCRT), vol. 9, pp. 5118-5122, Feb. 2021.

[2] Pathak, A. and Shetty, N.P., 2019. Indian stock market prediction using machine learning and sentiment analysis. In Computational Intelligence in Data Mining (pp. 595-603). Springer, Singapore.

[3] Reddy, V.K.S., 2018. Stock market prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 5(10), pp.1033-1035.

[4] Choi, H.K., 2018. Stock price correlation coefficient prediction with ARIMA-LSTM hybrid model. arXiv preprint arXiv:1808.01560.

[5] Wang, H., 2020, July. Stock price prediction based on machine learning approaches. In Proceedings of the 3rd International Conference on Data Science and Information Technology (pp. 1-5).

[6] Adhikar, A.J., Jadhav, A.K., KH, C.G.K. and HS, M.S.,2020 LITERATURE SURVEY ON STOCK PRICE PREDICTION USING MACHINE LEARNING.

[7] Ya Gao, Rong Wang, and Enmin Zou, "Stock Prediction Based on Optimized LSTM and GRU Models," Hindawi, vol. 2021, pp. 1-8, Sept. 2021.

[8] Babu, C.N. and Reddy, B.E., 2014, October. Selected Indian stock predictions using a hybrid ARIMA-GARCH model. In 2014 International Conference on Advances in Electronics Computers and Communications (pp. 1-6). IEEE

[9] Kadam, M.Y., Kulkarni, M.S., Lonsane, M.S. and Khandagale, A.S., A Survey on Stock Market Price Prediction System using Machine Learning Techniques.

[10] Thakkar, A., and Chaudhari, K., 2021. A comprehensive survey on deep neural networks for the stock market: The need, challenges, and future directions. Expert Systems with Applications, 177, p.114800.

[11] XIAOQIANG. (2019, April). What is a support vector machine? easyai.tech. Retrieved October 30, 2022, from https://easyai.tech/en/ai-definition/svm

[12] Jiayu QiuJ, Bin Wang, amp; Changjun Zhou. (2020). Structure of long short-term memory(Lstm). Forecasting stock prices with long-short term memory neural network based on attention mechanism. Retrieved October 30, 2022.

[13] Kostadinov, S. (2017). Gated Recurrent Unit. Understanding GRU Networks. Retrieved October 30, 2022, from https://medium.com/towards-data-science/understanding-gru-networks-2ef37df6c9be.

[14] thingSpeakRead.Mathworks. (n.d.). Retrieved October 30, 2022, from https://www.mathworks.com/help/thingspeak/calculate-simple-moving-average.html.

[15] Recurrent neural networks. Research Gate. (2019, February). Retrieved October 30, 2022

[16] Jakub Adamczyk . (2020, September 12). Make kNN 300 times faster than Scikit-learn's in 20 lines! towardsdatascience.com. Retrieved October 30, 2022, from https://towardsdatascience.com/make-knn-300-times-faster-than-scikit-learns-in-20-lines-5e29d74e76bb.

[17] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:14123555. 2014.

[18] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures. In: International Conference on Machine Learning PMLR. 2015;2342–50.

[19] Shewalkar A. Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. JArtif Intell Soft Comput Res. 2019;235–45.

[20] Vaishnavi Gururaj, Shriya V R and Dr. Ashwini K "Stock Market Prediction Using Linear Regression and Support Vector Machines" http://www.ripublication.com/

[21] K. Hiba Sadia, Aditya Sharma, Adarrsh Paul, SarmisthaPadhi, Saurav Sanyal "Stock Market Prediction Using Machine Learning Algorithms" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-4, April 2019

[22] C.H. Vanipriya1 and K. Thammi Reddy "indian Stock Market Predictor System"

[23] Edgar P. Torres P, Myriam Hernández-Álvarez,Edgar A. Torres Hernández, and Sang Guun Yoo"Stock Market Data Prediction Using Machine Learning Techniques

[24] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements"

[25] Mariam Moukalled, Wassim El-Hajj, Mohamad Jaber "Automated Stock Price Prediction Using Machine Learning"

[26] shita Parmar, Navanshu Agarwal, Sheirsh Saxena" Stock Market Prediction Using Machine Learning 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)

[27] Aparna Nayak, M. M. Manohara Pai and Radhika M. Pai" Prediction Models for Indian Stock Market"

[28] T. Manojlović and I. Štajduhar "PREDICTING "Stock Market Trends Using Random Forests: A Sample of The Zagreb Stock Exchange" MIPRO 2015, 25-29 May 2015, Opatija, Croatia

[29] Kuna Pahwa and Neha Agarwal "Stock Market Analysis Using Supervised Machine Learning"

[30] Radu Iacomin "Stock Market Prediction"2015 19th International Conference on System Theory, Control and Computing (ICSTCC), October 14-16, Cheile Gradistei, Romania

[31] Sumeet Sarode, Harsha G. Tolani, Prateek Kak, Lifna C S "Stock Price Prediction Using Machine Learning Techniques" International Conference on Intelligent Sustainable Systems (ICISS 2019)

[32] Karim, M. E., Foysal, M., Das, S. (2022, November). Stock Price Prediction Using Bi-LSTM and GRU-Based Hybrid Deep Learning Approach. In Proceedings of Third Doctoral Symposium on Computational Intelligence: DoSCI 2022 (pp. 701-711). Singapore: Springer Nature Singapore.

[33] Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena" Stock Market Prediction Using Machine Learning 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)

[34] Huynh, H. D., Dang, L. M., Duong, D. (2017, December). A new model for stock price movements prediction using deep neural network. In Proceedings of the 8th International Symposium on Information and Communication Technology (pp. 57-62).

[35] Ji, X., Wang, J., Yan, Z. (2021). A stock price prediction method based on deep learning technology. International Journal of Crowd Science, 5(1), 55-72.

[36] Saha, S., Singh, N., Mohan, B. R., Naik, N. (2021). A combined model of ARIMA-GRU to FORECAST stock price. In Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences: PCCDS 2020 (pp. 987-998). Springer Singapore.