International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Stock Prediction Using Machine Learning Methods

Er. Reshma Khan, Gagan Ajit Singh, Shubham Behal, Kartik Sharma, Saurabh Kumar
Dept. of Computer Science and Engineering,
Chandigarh University,
Mohali, Punjab, India

***Abstract*— The volatility and non-linear nature of the financial stock markets make it incredibly difficult to estimate stock market returns efficiently.. Investors need rapid access to precise information while trading stocks to make intelligent selections. Programable prediction approaches have demonstrated to be increasingly successful in forecasting stock prices with the introduction of artificial intelligence and better computing capacity. However, several variables impact the decision- making process as a stock market trades multiple stocks. Furthermore, it is impossible to forecast the behavior of stock prices. All of these elements make stock price prediction, both vital and tricky. This drives research into the most accurate prediction model that creates the fewest mistakes in its projections. This research studies machine learning approaches and algorithms in an effort to increase the accuracy of stock price prediction.**

***Keywords*— Machine Learning, Linear Regression, LSTM, SVM, Decision Tree, Random Forest**

I. INTRODUCTION

Stock market is a dynamic system based on non-linear, changing, and intricate behaviors. You may figure out the future value of business stock and other financial assets traded on an exchange by applying machine learning for stock price prediction.

A high amount of uncertainty, noise, non-stationarity, unstructured nature, and hidden linkages are features of the stock analysis discipline. A strategy for assessing or anticipating the probable price of a stock or other asset on stock exchange is stock price prediction. Predicting changes in the stock market price is exceedingly difficult due to a considerable lot of unpredictability in price volatility as well as a variety of other factors, such as political developments, broader economic conditions, and trader's expectations. In addition, the forecast incorporates additional elements that contribute to share prices' volatility and dynamic character, such as physical and psychological features as well as reasonable and irrational conduct. As a result, creating exact projections concerning stock values is tough. Programmers employ these tools to gather and educate the computer to detect a broad variety of complicated patterns.

II. OBJECTIVES

Using machine learning (ML), stock price prediction aims to provide precise and informed predictions about upcoming changes in financial markets. The main goals of this are:

1. *Risk Reduction:* To provide precise projections in order to minimise losses in volatile market conditions and assist traders and investors in identifying possible dangers related to certain stocks or portfolios.
2. *Improvement of Investment Decisions:* To provide investors with relevant information to help them decide whether to purchase, sell, or hold onto their stocks as well as to help them find the best times to enter and exit the market.
3. *Portfolio Management:* To assist investors select the optimal asset allocation using machine learning algorithms, so aiding in the creation of diversified portfolios that strike a balance between risk and return.
4. *Automated Trading:* To enable automated trading systems to execute buy and sell orders, automated trading systems will be able to make decisions fast and intelligently based on available market data.
5. *Trends to Expect:* To forecast both short- and long-term price patterns for stocks, enabling investors to profit on probable market movements before they happen.
6. *Enhancing Trading Techniques:* To improve trading methods' efficiency by adding predictive models, and to help trading professionals find the best entry and exit points as well as the right position sizing.
7. *Statistical Analysis:* To use quantitative models to complement existing fundamental and technical analysis techniques, and to uncover data correlations and patterns that may not be obvious through manual examination.
8. *Scenario Analysis and Stress Testing:* To assess the adaptability of investment portfolios to various market and economic conditions, and to predict outcomes by simulating different market situations using ML models.
9. *The Ability to Adapt to Shifting Market Conditions:* To design models that can endure a variety of market circumstances, including bull and bear markets as well as periods of excessive volatility.
10. *Including Outside Influences:* To take into account outside factors that can affect stock prices, such as geopolitical events, news emotion, and industry- specific variables. These inputs can be analysed by ML models to increase forecast precision.

III. METHODOLOGIES

The general mechanism for stock price prediction in this study is machine learning. This is an example of a data-driven strategy that uses computer algorithms to extract complex correlations, patterns, and trends from historical financial data. Through the utilization of machine learning models, this methodology enables the development of prediction systems that possess the ability to process and

comprehend extensive datasets, detect temporal relationships, and adjust to the dynamic and non-linear characteristics of financial markets. Enabling the models to identify small signals in the data so they can forecast future movements in stock prices is the main idea behind this technology. A range of models, from conventional linear regression to complex deep learning architectures, are covered by machine learning techniques, which provide a flexible toolkit. With a wide range of models available, it is possible to investigate the various aspects of stock price prediction with agility and to strive for precise projections in a constantly changing financial environment.

Three categories of machine learning implementations are distinguished by the kind of learning "signal" or "response" sent to a learning system.

I. Supervised Education

2. Unsupervised Education

III. Learning via Reinforcement.

I. Supervised Learning

A part of machine learning and artificial intelligence is popularly known as supervised learning. Its key characteristic is the way it trains computers to use tagged datasets to properly categorize data or forecast outcomes. The model changes its weight when input data is supplied into it during the cross-validation step until it has been fitted appropriately. Businesses may develop large-scale solutions for a large area of real-world based issues with the help of supervised learning, such as separating spam into a different folder from your email. The testing data is used to evaluate the algorithm's performance on unlabeled data. The loss function is used to test the accuracy of the algorithm and is changed until the error is adequately decreased. A model learns through time by feeding on data, a process known as supervised learning.

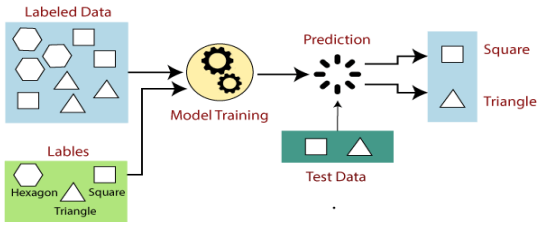


Fig. 1 Working behaviour of Supervised Learning

Supervised learning difficulties may be classified into two categories:

1. Classification: This approach reliably classifies test data into discrete categories. It tries to locate certain objects within the collection and makes deductions regarding their appropriate labelling or description. Random forests, KNN, Decision trees, Support evctor machines (SVM), and linear classifiers are a few of the most often used classification algorithms.
2. Regression: This technique establishes the correlation between the independent and dependent variables. It's often used to create estimates, such as sales income predictions for businesses. Common regression approaches include logistic, polynomial, and linear regression algorithms.

II. Unsupervised Learning

Machine learning techniques are used in unsupervised learning to evaluate and classify unlabelled samples. These algorithms can locate data clusters or hidden patterns without requiring human assistance. Because of its capacity to spot patterns in data, it is perfect for consumer segmentation, picture recognition, exploratory data analysis, and cross-selling tactics. Unsupervised learning is used to analyse unlabelled and uncategorized data in order to find latent structures. The data scientists utilise the training datasets to train the algorithms, which initiates the unsupervised learning process. These databases include data points that are not tagged or classified. Finding patterns in the dataset and categorising each individual data point in accordance with those patterns is the algorithm's learning goal.

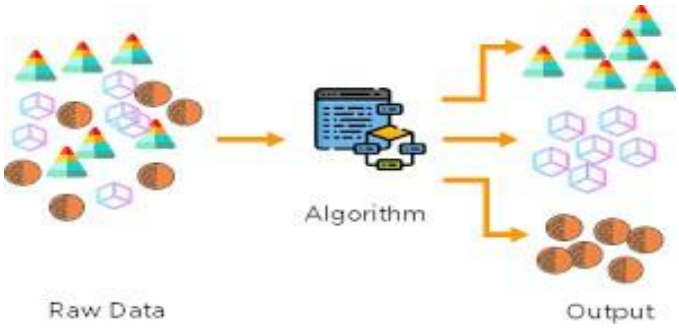


Fig.2 Working behaviour of Unsupervised Learning

There are two sorts of problems that can be solved using the unsupervised learning algorithm:

1) Clustering: Using clustering, items are arranged into groups according to how similar they are, with those that are less or non-existent remaining in one group. Through the use of cluster analysis, similarity between data objects is found and their existence or absence is categorized.

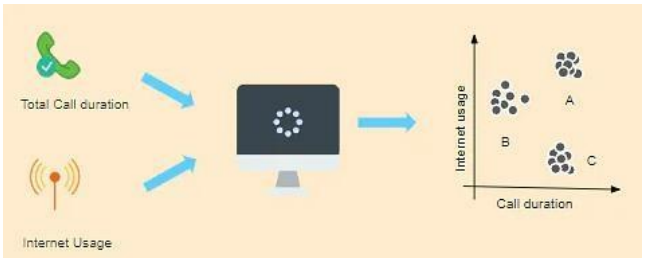


Fig. 3 Clustering – Unsupervised Learning

2) Association: To discover relationships between variables in a large database, association rules are an unsupervised learning approach. It identifies the group of items that are present together in the dataset. The efficacy of marketing strategies is increased by the association rule. People who purchase X (bread, for instance) are more inclined to purchase Y (butter/jam). An association rule may be helpfully shown with the use of market basket analysis. identifying the items that were purchased simultaneously.

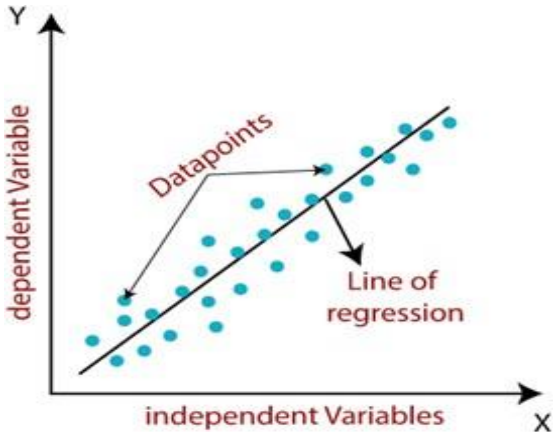


Fig.4 Association - Unsupervised Learning

III. Reinforcement Learning

By incorporating feedback from its own experiences and actions, Reinforcement Learning (RL) is a machine learning way that enables an agent or user to learn via trial and error in an interactive or user-friendly surroundings. In reinforcement learning, incentives and punishments act as markers of acceptable and wrong conduct. In this scenario, the aim is to determine the appropriate action model to maximize rates in the action-reward feedback loop of a generic reinforcement learning (RL) model.

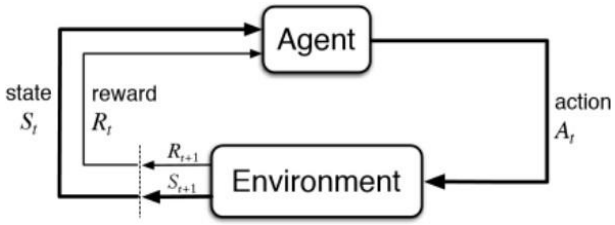


Fig.5 Basic RL model

IV. ML TECHNIQUES FOR STOCK PREDICTION

I. Linear Regression

Linear regression is a fundamental and interpretable model in the realm of stock price prediction. It presupposes that there exists a linear relationship between the predictor variables, i.e., the characteristics and the target variable, i.e., the stock price.

It aims to discover the best-fitting line, also known as a hyperplane, that minimizes the sum of squared differences in the midpoint of the predicted and actual values by modeling the relationship between the input data and the goal variable as a linear equation.

It aims to generate a linear equation that can explain or predict changes in stock price by estimating the coefficients of these features. It provides insights into the significance of features by making it evident whether aspects are positively or negatively connected with price fluctuations In Linear Regression we collect historical stock price data for various companies across different industries. This dataset includes features such as daily opening and closing prices, trade volume and market sentiment indicators and several.

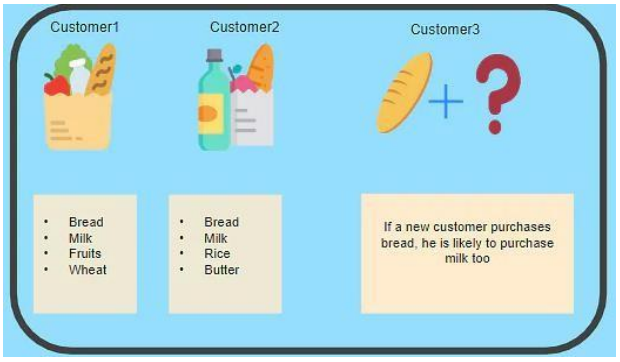


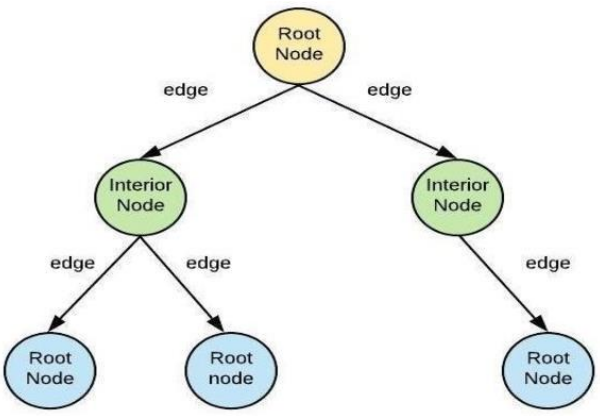
Fig. 5 Graph of Linear Regression

other market-related parameters as predictor variables. To use linear regression, we regarded the historical stock price as the objective variable and applied numerous market-relevant parameters as predictor variables. We separated the data into training and testing sets to evaluate the model's performance. The success of linear regression in stock price prediction is dependent on the underlying data and the characteristics used for the model. Linear regression can produce robust forecasts when stock prices have significant linear correlations with relevant inputs. However, it may fall short in effectively depicting the underlying dynamics in scenarios with complicated, non-linear connections or quick market fluctuations.

II. Decision Tree

It is a supervised learning approach. It is constructed of a structure that resembles a tree, with each node reflecting a choice made in response to a characteristic and each branch representing a potential conclusion. Using a decision tree, one may traverse the tree from the root to a leaf node to make decisions or anticipate results. One of the fundamental downsides of decision trees is that, in the absence of adequate restrictions, they may become overfit. When a model is extremely complicated, it might match the noise in the data rather than the underlying trend, which is known as overfitting. When the model is employed with fresh data, this might result in mediocre performance.

Fig. 6 Structure of Decision Tree



III. Artificial Neural Networks (ANN)

ANN is a technique used in artificial intelligence to educate computers to interpret data similarly to how the human brain does. A type of machine learning called deep learning uses linked nodes or neurons arranged in layers to resemble the organisation of the human brain. It creates an adaptable framework so that computers can grow and learn from their errors. ANN aims to do increasingly precise tasks that are complicated, including facial recognition or document summarization. These systems will imitate how people learn because they are based on how brain neuron's function. Programmers utilise these tools to collect and train the machine to recognise a wide variety of intricate patterns. To put it simply, it is a collection of algorithms that mimic the functioning of the human brain in order to identify the fundamental relationship among groups of the data.

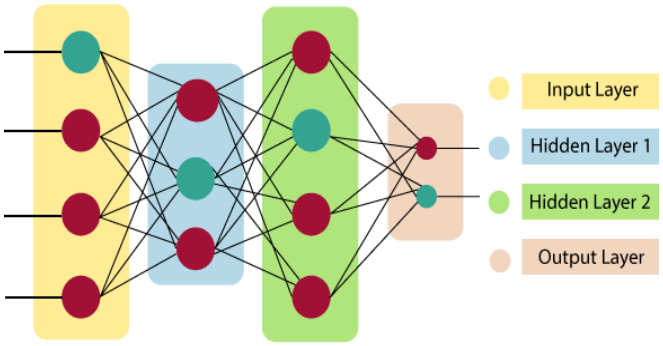


Fig. 7 Structure of ANN

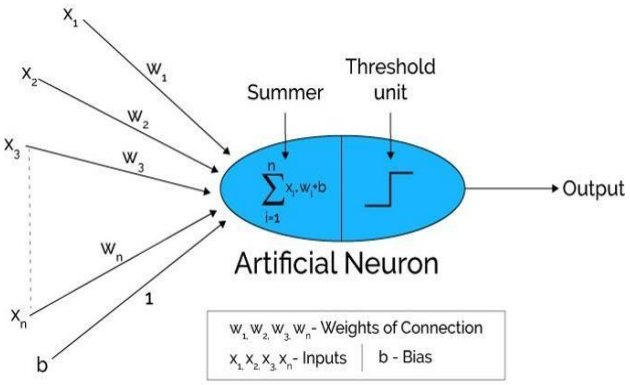


Fig. 8 Working of ANN

IV. Random Forest

Random Forest applies to a range of decision trees on various parts of the given input data to increase projected accuracy or efficiency of the dataset. Using the predictions from every decision tree in addition to a majority voting method, the random forest predicts the result rather than relying just on one decision tree. Compared to other algorithms, the Random Forest approach takes shorter training time. It functions well and generates output predictions with exceptional accuracy, even with the large dataset. Even in cases where a substantial quantity of data is absent, it can nevertheless be accurate. Certain decision trees may properly forecast the dataset's class while others may not since the random forest mixes several trees to accomplish so. Collectively, nevertheless, every tree forecasts the right result. For a better Random Forest classifier, the following two presumptions are made:

1. In order for the classifier to accurately predict results instead of generating conjectures, the feature variable of the dataset needs to have some real values.
2. Very low correlations are required between the predictions made by each tree.

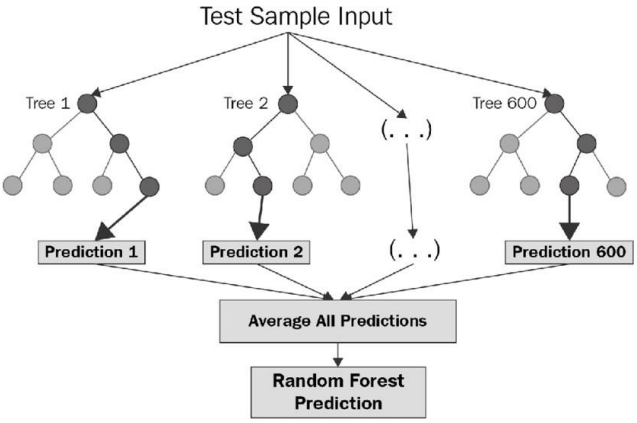


Fig. 9 Structure of Random Forest

V. Support Vector Machines (SVM)

SVM can be used to address problems with both regression and classification. Selecting the optimal N-dimensional space hyperplane for data point classification into various feature space classes is the primary goal of the SVM technique. When data points cannot be separated linearly, SVM employs a high-dimensional feature space mapping to categorise them. When a separator is found between the formation is processed to provide a hyperplane representation of the separator. The extreme vectors and points are picked by SVM to build the hyperplane.

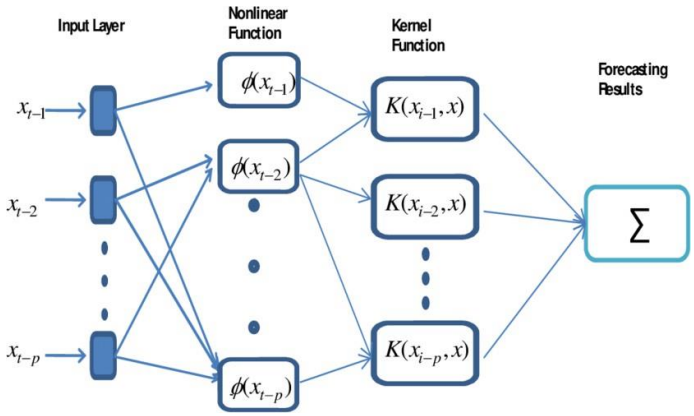


Fig. 10 Structure of SVM

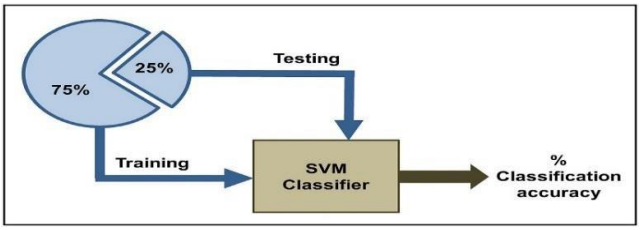


Fig. 11 Learning Architecture of SVM

VI. Long Short-Term Memory Networks (LSTM)'

Sequential neural networks that preserve information and are capable of deep learning are called LSTM. It's a sort of recurrent neural network that can address the vanishing gradient problem that RNNs have. Hochreiter and Schmid Huber created LSTM to tackle a problem with standard running and machine learning approaches. LSTM is implemented in Python using Kera's module. Assume that you are seeing a film and recall the previous scene, or that you are reading a book and know what happened in the prior chapter. In a manner similar to this, RNNs retrieve and use prior knowledge to analyse the present input. The RNN is unable to recollect long-term dependencies due to the reduction gradient. LSTM deliberately avoids long-term dependency concerns in its architecture.

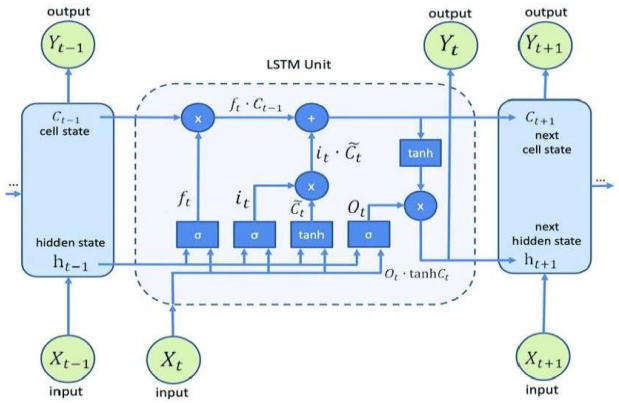


Fig. 12 Structure of LSTM

VII. Convulational Neural Networks (CNN)

The billions of neurons that make up CNN are grouped a certain way. The design of CNN is comparable to the connection pattern of the human brain. The structure of the neurons in a CNN is exactly like that of the brain's frontal lobe, which handles the processing of visual inputs. By ensuring that the full visual field is covered, this structure helps neural networks avoid the piecemeal image processing issues that arise when they are fed pictures in low-resolution bits. CNN uses parameter sharing; every layer node in CNN is connected to every other layer node. CNN is also associated with a weight; when the layer filters pass over the picture, the weights stay the same and a phenomenon called parameter sharing takes place. As a result, the CNN system experiences less compute stress.

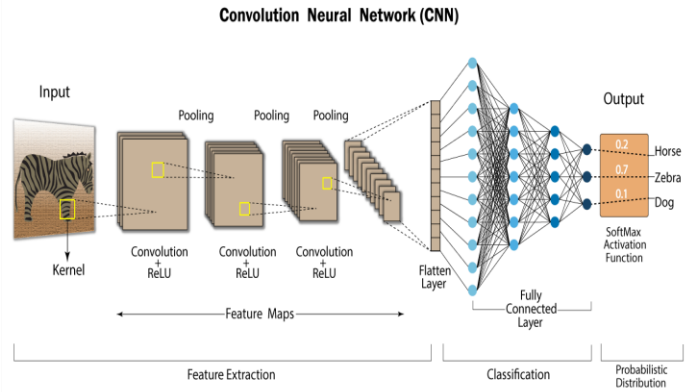
CNN consists of three layers:

Convolutional Layer: Convolutional layers are the core building blocks of neural networks (NNNs), where most computations occur. Convolution is a process in which a kernel or filter inside the layer traverses the receptive fields of the picture and looks for features. The kernel repeatedly iteratively scans the entire image. Every iteration result in the calculation of a dot product between the input pixels and the filter. The result of the dots is a convolved feature, also known as a feature map. Ultimately, the image is converted into numerical values in this layer so that the CNN can comprehend it and identify patterns that are valuable.

1. Pooling Layer: The pooling layer gives the input a kernel or filter. The pooling layer also loses some information throughout the process of lowering the input's parameter count. On the other hand, this layer simplifies and improves the performance of CNN.
2. Fully Connected Layer: The features obtained from the upper levels are used for classification in the CNN's FC layer. All the nodes or activation units in the layer below are linked to all the inputs or nodes in the layer above. All of the layers are not connected in the CNN, though, as this would result to an extremely coupled network. In addition, it would increase losses, be computationally expensive, and produce output of

lower quality.

Fig. 13 Structure of CNN



VIII. Recurrent Neural Networks (RNN)

RNNs operate by reusing the output of one layer as an input and storing it to forecast the output of the following layer. An RNN is one sort of neural network that can imitate sequence data. RNNs are obtained from feed-forward networks and behave similar to human brains. Because recurrent neural networks can predict sequential data, they have an edge over other approaches. Applications of RNN include voice recognition, speech recognition, time series prediction, and natural language processing (NLP).

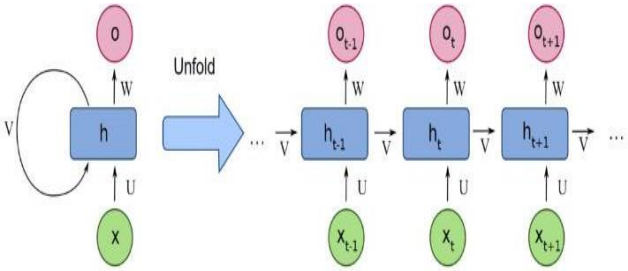


Fig. 14 Structure of RNN

IV. CHALLENGES

Stock price prediction is a difficult endeavour with several unique difficulties.
The following are some of the main issues with stock price prediction:

1. *Market noise and randomness:* Numerous factors, including news stories, economic data, investor mood, and more, can affect stock prices. It might be hard to identify significant patterns in the midst of all of this noise.
2. *Non-Stationary Data:* Non-stationary behaviour is exhibited by stock prices, where statistics' mean and variance change over time. Using traditional time-series analysis techniques becomes challenging as a result.
3. *Data Availability and Quality Challenges:* The quantity and quality of data can differ substantially. For both historical and real-time data, this can lead to bias in model
4. *Model complexity and overfitting:* Complex models might detect false patterns in the data rather than noise. Overfitting could ensue from this, where a model performs well on historical data but badly on new, unanticipated data.
5. *Lack of Causality:* Correlation does not always imply causation. It is not always the case that two variables that are correlated are causally related. Understanding the causal relationships in the financial markets is challenging.
6. *Long-term Trends and Structural Changes:* Recognising and accounting for long-term trends and structural changes in the market can be challenging. For example, it could be difficult to predict changes in market dynamics brought on by technological advancements or governmental reforms.
7. *Efficiency of the Stock Market:* Stock prices are assumed to reflect all available information under the premise of efficient markets. In actuality, markets could not always be completely efficient, and there might be chances to make money through predictions.

V. CONCLUSION

The integration of machine learning algorithms with sentiment analysis highlights the critical function of data-driven approaches in strengthening stock price prediction and facilitating more prudent investing choices. These state-of-the-art methods, such Random Forest and Support Vector Machine (SVM), demonstrate how well they may improve forecast accuracy and promote a thorough comprehension of the complex stock market environment. The incorporation of data-centric methodologies is expected to transform stock market analysis as financial markets develop. This will enable investors and stockbrokers to utilise advanced tools for analysing market patterns, seizing opportunities, and managing associated risks.

REFERENCES

[1] W. Huang et al., “Forecasting stock market movement direction with support vector machine,” Computers & Operations Research, 32, pp. 2513–2522005, 2005
[2] Vatsal H. Shah, “Machine learning techniques for stock prediction,”
[3] J. Moody, et al., “Learning to trade via direct reinforcement,” IEEE Transactions on Neural Networks, vol. 12, no. 4, Jul. 2001.
[4] Shen, Shunrong, Haomiao Jiang, and Tongda Zhang. "Stock market forecasting using machine learning algorithms." (2012).
[5] Lin P, Su S, Lee T. Support vector regression performance analysis and systematic parameter selection [J]. IEEE, 2005
[6] Olaniyi, S.A.S. & S., Adewole & Jimoh, R., (2011). Stock trend prediction using regression analysis-A data mining approach.
ARPN Journal of Systems and Software
[7] Rinehart, M. (2003). Overview of Regression Trend Channel (R.T.C.)
[8] Kyoung-jae Kim (2003). Financial time series forecasting using support vector machines, neurocomputing
[9] Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., & Zhang, J., (1998). Daily stock market forecast from textual web data. In Systems, Man, and Cybernetics
[10] Jasic, T., & Wood, D. (2004). The profitability of daily stock market indices trades based on neural network predictions: Case study for the S&P 500, the DAX, the TOPIX and the FTSE in the period 1965–1999. Applied Financial Economics
[11] Kim, H. J., & Shin, K. S. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. Applied Soft Computing
[12] Kim, K. J., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert systems with Applications
[13] Kim, K. J., & Lee, W. B. (2004). Stock market prediction using artificial neural networks with optimal feature transformation.
Neural computing & applications
[14] Kim, M. J., Min, S. H., & Han, I. (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. Expert Systems with Applications
[15] Kumar, L., Pandey, A., Srivastava, S., & Darbari, M. (2011). A hybrid machine learning system for stock market forecasting. Journal of International Technology and Information Management
[16] Chavan, P. S., & Patil, S. T. (2013). Parameters for stock market prediction. International Journal of Computer Technology and Applications