Study and Analysis of Stock Market Prediction Techniques

Siddhesh Kokare[1], *Anvit* Kamble[2], *Shubham* Kurade[3] *and Deepali* Patil[4]

[1]Department of Information Technology, Ramrao Adik Institute of Technology, Navi Mumbai, India siddheshkoare512@gmil.com

[2]Department of Information Technology, Ramrao Adik Institute of Technology, Navi Mumbai, India anvitkamble@gmail.com

[3]Department of Information Technology, Ramrao Adik Institute of Technology, Navi Mumbai, India shubhamkurade21@gmail.com

[4]Department of Information Technology, RAIT,D.Y. Patil Deemed to be University, Navi Mumbai, India deepali.patil@rait.ac.in

Abstract. Stock marketplace is a complicated and demanding system in which people make more money or lose their entire savings. The stock market prediction having high accuracy yields more profit for stock investors. Stock market data is generated in a very large amount and it varies quickly every second. The decision making in stock marketplace is a very challenging and strenuous task of financial stock market. The development of efficient models for prediction decisions is very difficult because of the convolution of stock market financial data and should have high accuracy. This study attempts to compare existing models for the stock market. Various Machine learning methods like Long Short Term Memory (LSTM), Convolution Neural Networks (CNN) and Convolution Neural Networks – Long Term Short Memory (CNN-LSTM) have been used for the comparison. The models are estimated using conventional strategic measure: MAE (Mean Absolute Error). The measured low values indicates that the models are effective in predicting stock prices.

Keywords: Stock Market Prediction, LSTM, CNN, CNN-LSTM.

1. Introduction

Stock market forecasts are very crucial as it is utilized by many business people and general public. People will both benefit cash or lose their whole existence financial savings in stock market. Building precise models is tough as it relies upon more than one element along with news, social media information, fundamentals, manufacturing of the company, authorities bonds and country' s economics. Prediction model which consists only one factor might not be precise. There are several theories regarding stock market that have been conceptualizing over the years. Those theories try to expalin whether the market can be beaten or try to explain stock market nature. The market price of the stock integrates all the information about that stock in that particular time frame. The changing tendency of the stock prices has been regularly recognized as a total hassle inside the financial field [1]. Stock prices are suffering from diverse inner and outside factors, which includes home and overseas financial environment, global situation, enterprise prospect, economic information of indexed companies, and stock market operation [2, 3].

The conventional analysis is primarily based on finance and economics which uses basic and technical analytical methods. Firstly, the fundamental analysis focuses on the inherent stock values and qualitatively analyses the external factors like interest rates, exchanged rates, inflation, industrial policy, listed companies finances and international relations, etc. which affects the stocks. Secondly, the technical analysis mostly focuses on the stock price direction, trading volumes and psychological expectation of investors that focuses mostly on stock market by using tools like k-line chart or by analyzing individual stock directions in the stock indexes. The above methods are stilled the most generally used methods for many companies and investors [4,5].

Traditional fundamental analysis accuracy is tough to be convincing because the prediction results are highly dependent on the professional quality of the analytics and the influential factors are in a long term cycle. The stock data have the traits of random walk in a financial time series. The accuracy of usage of only time series model is questioned due to unknown and high noise characteristics of the financial time series [6].

There are certain limitations on clearly predicting stock price trends when using linear time series forecast version or the neural network version. Currently, by combining the benefits of different methods to enhance the hybrid approach is now an economic improvement trend for time series deep learning [7].

Therefore, in order to make the better use of the time series, thorough investigation of the characteristics of the records and accuracy improvement of the stock price forecasting can be done. This paper compares stock prices forecasting approaches based primarily on CNN, LSTM and CNN-LSTM.

2. Literature Survey

The financial marketplace is noisy, non-parametric dynamic and there are mainly two types of forecasting techniques: Technical analysis technique and machine learning techniques [8]. The conventional econometric techniques or equations with parameters aren't appropriate for studying complicated large dimensional and noisy financial data. In the paper proposed by Aparna et al. [9], consideration of various parameters of various datasets as done, it was observed that Decision Boosted Tree was performing better when compared to SVM and logistic regression. The paper proposed by Vijh et al. [10], dataset of five companies from 2009-2019 having new parameters for better prediction, such as High - Low , Open-Close, 7 day average stock price, 14 day average stock price, 21 days average stock price , last 7 days standard deviation was used. Comparative analysis based on RMSE, MAPE and MBE results clearly shows that ANN provides better stock prediction when compared to RF. The paper proposed by Hyeong et al.[11], the model performance was validated on both different time periods with several metrics like MSE, MAE and RMSE. By analyzing the testing results it was observed that Arima-Lstm hybid performs far better when compared to other financial models. The paper proposed by Pushpendu et al.[12] it mainly focuses on application of Random Forest and LSTM to predict stock prices directional movements. It was observed that the LSTM outperforms random forests. The paper proposed by Mehtabhorn et al. [13] it basically compares the various types of machine learning techniques and algorithm which is used in finance and stock market prediction. The paper proposed by Wenjie Lu et al. [14], The CNN-LSTM model is used to predict the closing price of a stock price the next day. Experimental results show that CNN-LSTM have highest accuracy and best performance compared to CNN, RNN, LSTM< MLP and CNN_RNN. The paper proposed by Nusrat Rouf et al. [15] comparisons of ANN, SVM, NB and DNN was carried out. SVM was the most popular technique used for SMP. It was observed that ANN and DNN performs more accurate and provides faster prediction. The paper proposed by Jingyi Shen et al. [16] used a comprehensive deep learning system. Prediction was carried on the datasets of Chinese stock market using the LSTM models. It was observed that the LSTM model achieved high prediction accuracy and outperformed the major models. The paper proposed by D. Wei et al. [17], the prediction was performed on the datasets using various LSTM models. It was observed that Vanilla LSTM, Stacked LSTM and Bidirectional LSTM are the commonly used LSTM models. BI-LSTM was having greater accuracy and low error when compared to other models. The Paper proposed by Sheng Chen and Hongxiang He et al. [18], CNN model was used for making Stock prediction which was perform using conv 1d function to process 1d data in convolution layer. It was observed that if source data is sequential then the model is efficient and can even be used to make predictions. The paper proposed by Wu et al. [19] with leading indicators prediction was performed on dataset using hybrid CNN-LSTM model. It was observed that CNN-LSTM model was achieving greater accuracy when compared with CNN and LSTM models. The paper proposed by Xuan Ji et al. [20], MAE, RMSE and R-square values are calculated to evaluate the performance of various prediction models. It was observed that CNN-LSTM model out performs well when applied on various stock prices. The paper proposed by Vanukuru, Kranthi et al. [21], the SVM model was used for predicting the stock index movements. It was observed that model generates higher profit as compared to selected benchmarks. The paper proposed by A M Pranav et al. [22], the sentimental analysis was performed on stock prices to forecast stock price variations. It was observed that machine learning models were performing well on various datasets

3. Comparisons of Existing Machine Learning Models

3.1 Convolution Neural Network (CNN):

Sheng Chen [16], proposed a CNN model for creating stock prediction that use the conv1d function to process the 1D data in the convolution layer. CNN is a feedforward neural network that performs very well in image processing and natural language processing. If implemented correctly, it can even predict forecasting of the time series. The local perception and weight distribution of the CNN can significantly reduce parameter range thereby improving the performance of model learning.

The CNN as shown in Figure 1, particularly consists of a convolution layer as well as the pooling layer. Each convolution layer consist of various convolution kernel and its formula is shown in equation (1).

$$o_v = \tanh(v_i * k_w + v_b) \qquad (1)$$

where o_v is the output value after convolution, tanh is the activation function, v_i is the input vector, k_w is the convolution kernel weight, and v_b is the convolution kernel bias.

The CNN model extracts the features map with varying details across convolution layers of stock data. The stock data includes stock market performance of assets over the period of IPO (initial public offering - private companies offers its share to public in new stock issuance) introduction to current date. This is inherently temporally interdependent data which has been discovered in the EDA (Exploratory Data Analysis – the process of inspection of the dataset to find patterns, irregularity and structure hypothesis based on the comprehension of the stock dataset) phase. This temporal dependency is extracted as a 2D feature map by CNN which in turn is passed through dense layers to generate single continuous output that is target variable which is open price of the stock.

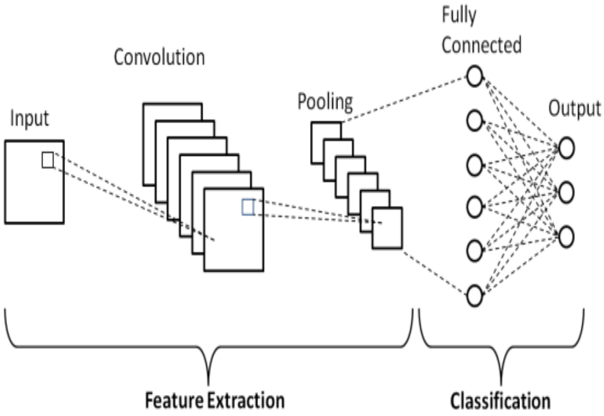

Figure 1. Convolution Neural Network (CNN)

3.2 Long Short term Memory (LSTM):

Xuan Ji [20], developed a new stock price forecasting model based on deep learning technology that uses Doc2Vec, SAE, wavelet transform, and LSTM mode. It mainly focuses on feature selection of stock financial features and text features (through social media like investors' comments and news published regarding stocks published by the companies).

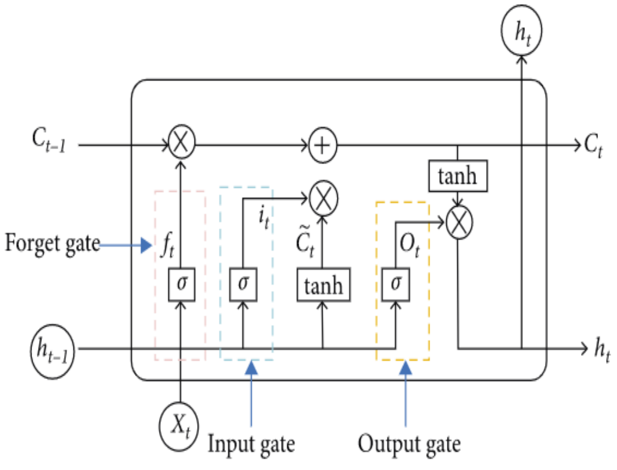

Figure 2. Long Short term Memory (LSTM) [14]

As shown in Figure 2, the LSTM memory cells consists of three parts: a forget gate, an input gate, and output gate. LSTM is used in analytical approach because LSTMs can store important information in the past and forget about other information. The input gate adds cell state information, the forget gate removes the information that is not needed in the model and the current gate selects the information that is displayed as an output. LSTMs combines the result of previous gate and the current gate and further, predicts the next state using the gain correlation.

LSTM is suited to extract long term dependencies in sequential data with temporal dependencies as it eliminates vanishing and exploding gradient in the backpropagation phase. The temporal dependency of stock market data is modelled by unfolding the data in time and passing through LSTM which predicts the output at next time step. The last price, volume and date provides input to the model and the open price is produced as the output of the target variable.

3.3 Convolution – Long Short term Memory Hybrid (CNN-LSTM):

The paper proposed by Wenjie Lu [14], states that a CNN-LSTM model is used to predict the closing price of the stocks of the next day. This method takes opening price, highest price, lowest price, closing price, volume, turnover, ups and downs, and changes in stock data as inputs, uses CNN to characterize the input data, and uses LSTM output to extract. It then learns the characteristic data and predict the closing price of the stock price the next day.

In CNN, it has the notions of listening to the maximum apparent features within side the line of sight, so its miles is extensively utilized in feature engineering. LSTM has the property of increasing overtime and is widely used in time series. The model structure diagram is shown in Figure 3. The CNN captures the spatial dependencies in the stock data inherent to the images. LSTM resolves the issue of vanishing or exploding gradient associated to the long term temporally dependent stock data. The combination of CNN and LSTM is tested in the predictive model. The CNN layers are used as initial layers which extracts the features in the sequential stock data and LSTM is then cascaded to incorporate long term dependency preservation in the features extracted by the CNN layers. At last, fully connected layers have been added to give the single continuous result.

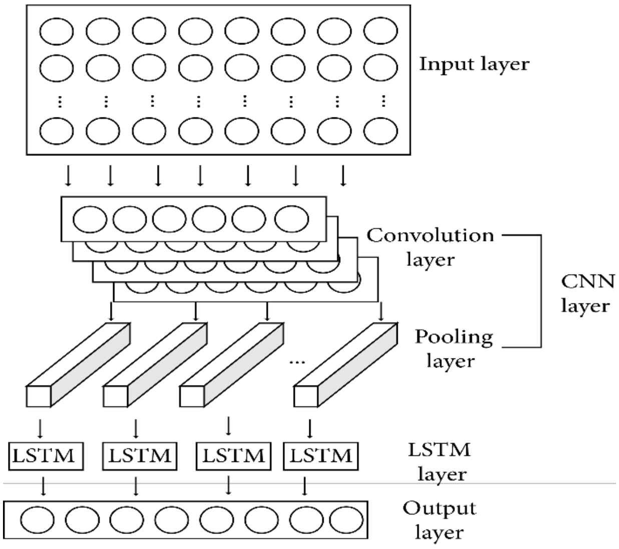

Figure 3. Convolution – Long Short term Memory Hybrid (CNN-LSTM) [14]

The models analyses the dataset in order to remove null values in the columns or replace them with mean or median values and also to detect the relationship between the parameters to determining the important parameters that affects the stock prices. The models handles the categorical and date type data by identifying the columns/attributes datatypes. The date data type is handle by diffusing it into three components. This generates the report regarding the dataset which is returned to the models which is used for further processing. The dataset is divided into training dataset (80% of original dataset) and testing dataset (20% of original dataset). The training of the models is done on the training datasets whose performances (models predicted values on testing dataset) are compared to the testing datasets actual values. This is done by the models which returns the performance metric calculated on the testing dataset. The models then gives the output of the performance report of all the three models.

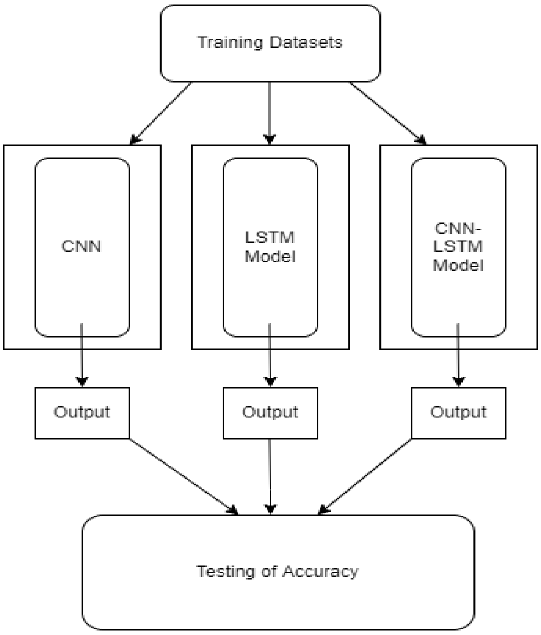

Figure 4. Comparison of CNN, LSTM, CNN- LSTM models

4. Results and Discussion

In the proposed method diverse recognized datasets are used which are TATASTEEL, TATAMOTORS, VEDL, BHARTIAIRTEL and ITC etc. These companies dataset are decided from various diverse sectors such as oil, telecommunication and from many others. The datasets are accumulated from 2000-2021 for evaluation.
The models are evaluated for TATASTEEL, TATAMOTORS, VEDL, BHARTIARTL and ITC dataset stocks.

The parameters used in the stock dataset are date, previous closed, high, low, last, close, volume and turnover. The date is when a stock is purchased/sell or when a stock transaction takes place. The previous closed is the previous closed price of the stock. The open price is the stock price when opened in a stock market on a particular date. The high is the highest stock price on a particular date. The low is the lowest stock price on a particular date. The last is the last price occurred for the last trade of a day. The close is the closing price of a stock is a price at which the stock closes at end of the trading hours in the stock market. The volume indicates how many are sold and bought in a given time duration.

The parameters settings for CNN model are convolution layer filters are 128, convolution layer kernel size is 3, convolution layer padding is same, learning rate is 0.001, loss function is mean absolute error, epochs is 25 and activation function is relu. The parameters settings for LSTM model are activation function is relu, learning rate is 0.001, loss function is mean absolute error, epochs is 15 and dropout is 0.2. The parameters settings for CNN-LSTM are convolution filters are 64, convolution layer kernel size is 3, convolution layer padding is same, batch size is 4, activation function relu, learning rate is 0.001, optimizer is adam, epochs is 15 and loss function is mean absolute error.

The dataset is partitioned into 80% training dataset and 20% testing dataset. The mean absolute error (MAE) is calculated for the evaluation of the forecasting effect on CNN, LSTM and CNN-LSTM models. The MAE equation (2) is given below,

$$MAE = \frac{\sum_{i=1}^{n} |p_v - t_v|}{n} \qquad (2)$$

in which t_v is true value and p_v is predictive value. Smaller the MAE value, the higher the accuracy of the models. The ML models in this work predict the open price of the several stocks on the nth day, in which n is the pre-decided value. The analysis of the models shows an overall average error (MAE) of 0.17586 for CNN-LSTM, 0.21204 for LSTM, and 0.22982 for CNN.

The graphs shown in Fig.5, Fig.7, and Fig.9 depict the MAE result for open price of TATASTEEL of CNN-LSTM, LSTM and CNN respectively. Similarly, the graphs shown in Fig. 6, Fig. 8 and Fig. 10 depict the loss for open price of TATASTEEL of CNN-LSTM, LSTM and CNN respectively.

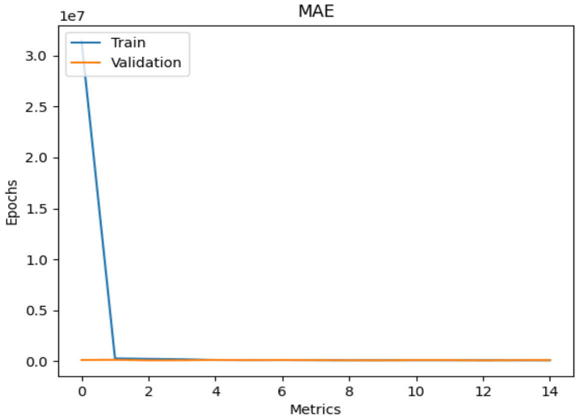

Figure 5. CNN-LSTM MAE

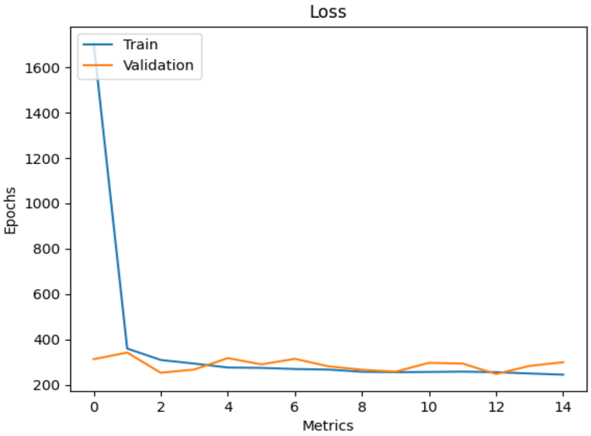

Figure 6. CNN-LSTM Loss

The forecasting of the open price of TATASTEEL using CNN-LSTM model shows that its results are the best among the three models.

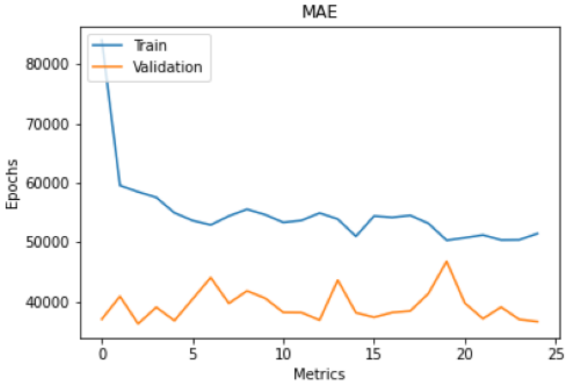

Figure 7. LSTM MAE

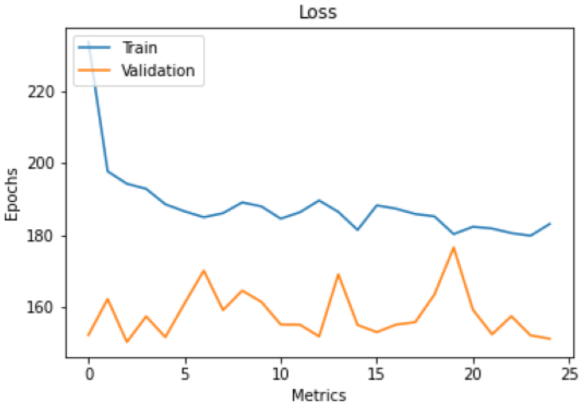

Figure 8. LSTM Loss

The forecasting of the open price of TATASTEEL using LSTM models shows that its results are better than CNN but lags behind.

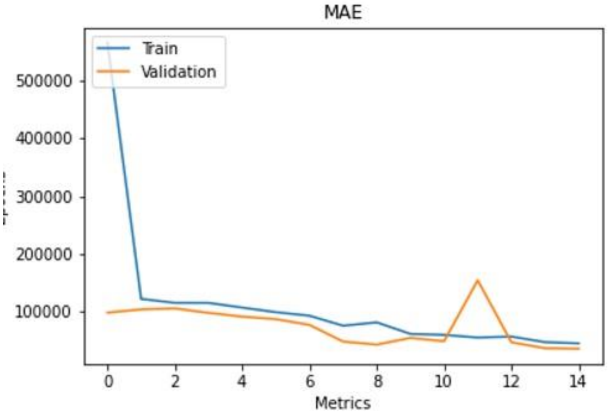

Figure 9. CNN MAE

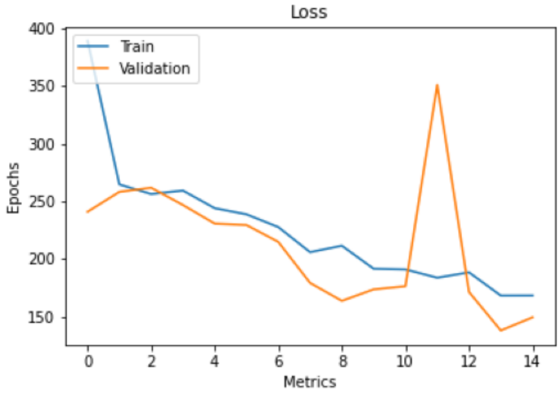

Figure 10. CNN Loss

The forecasting of the open price of TATASTEEL using CNN model shows that its lags behind the CNN-LSTM and LSTM models.

From Figures 5-10, the MAE and the loss graphs are generated using CNN-LSTM, LSTM and CNN respectively for the stocks datasets. The CNN-LSTM shows overall least variation for the datasets and hence CNN-LSTM has the highest forecasting accuracy in terms of forecasting performance.

Table 1. Comparison of Algorithms

MAE (Mean Absolute Error Values)				
Sr No.	Stocks	CNN	LSTM	CNN-LSTM
1	TATASTEEL	0.2838	0.2373	0.2051
2	TATAMOTORS	0.1878	0.1707	0.1351
3	VEDL	0.0154	0.0151	0.0125
4	BHARTIARTL	0.3092	0.2892	0.211
5	ITC	0.3529	0.3479	0.3156
		0.22982	0.21204	0.17586

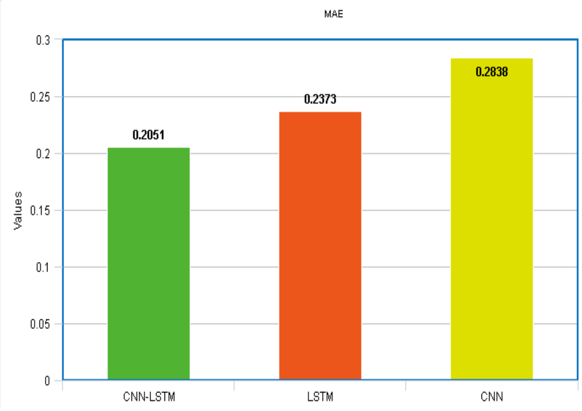

Figure 11. MAE comparison on TATSTEEL

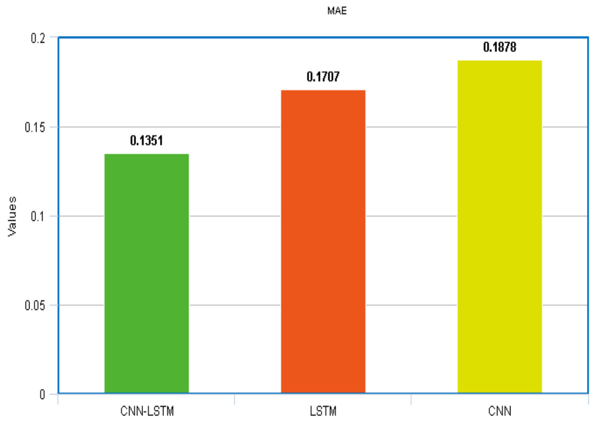

Figure 12. MAE comparison on TATAMOTORS

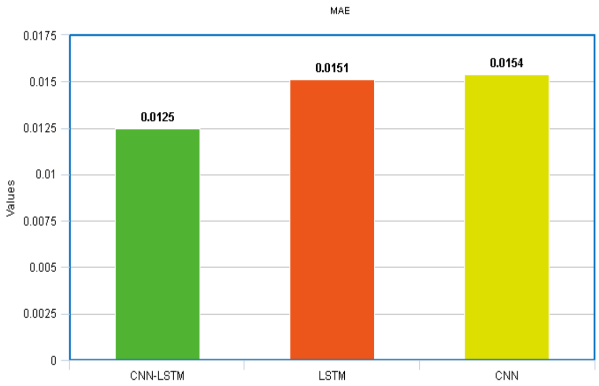

Figure 13. MAE comparison on VEDL

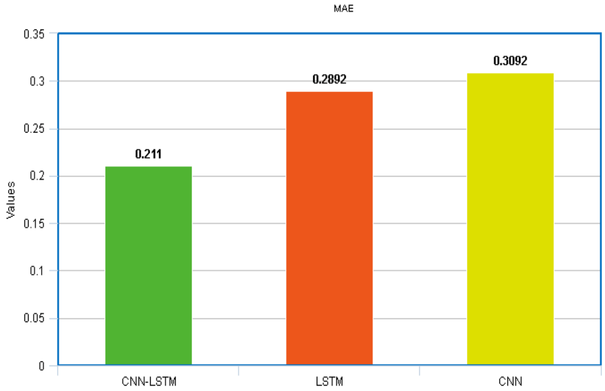

Figure 14. MAE comparison on BHARTIARTL

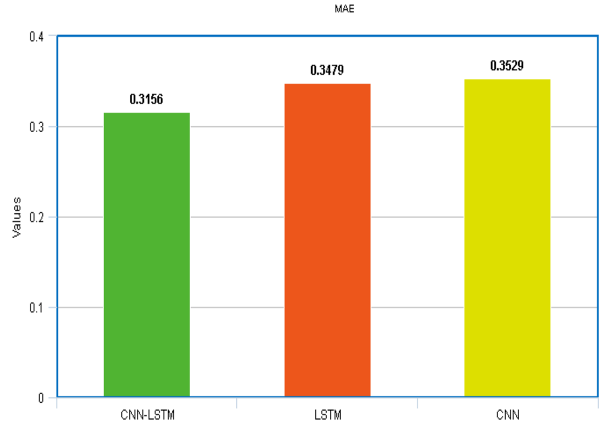

Figure 15. MAE comparison on ITC

In the Figures 10-15, the MAE values of CNN-LSTM, LSTM and CNN are calculated and their bar graphs have been plotted of the stock datasets. It is observed that CNN has the highest MAE values while CNN-LSTM has the lowest MAE values. Thus, this concludes that the CNN-LSTM Hybrid Model has the most smaller MAE values among the three models.

5. Conclusion

This report proposes a CNN, LSTM and CNN-LSTM based models to forecast the stock prices data according to the sequential traits of the stock prices data. The models uses stock datasets parameters such as close price, open price, high price, low price, previous closed price, turnover and volume. These parameters are used as the input to the models for training and testing for stock price prediction. The experimental results shows that in most of the cases, the CNN-LSTM performs very much better than CNN and LSTM models.

References

1. R. Vanaga and B. Sloka, “Financial and capital market commission financing: aspects and challenges,” Journal of Logistics, Informatics and Service Science, vol. 7, no. 1, pp. 17–30, 2020.
2. L. Zhang and H. Kim, “-e influence of financial service characteristics on use intention through customer satisfaction with mobile fintech,” Journal of System and Management Sciences, vol. 10, no. 2, pp. 82–94, 2020.
3. L. Badea, V. Ionescu, and A.-A. Guzun, “What is the causal relationship between stoxx europe 600 sectors? But between large firms and small firms?” Economic Computation And Economic Cybernetics Studies And Research, vol. 53, no. 3, pp. 5–20, 2019.
4. J. Sousa, J. Montevechi, and R. Miranda, “Economic lot-size using machine learning, parallelism, metaheuristic and simulation,” Journal of Logistics, Informatics and Service Science, vol. 18, no. 2, pp. 205–216, 2019.
5. A. Coser, M. M. Maer-Matei, and C. Albu, “Predictive models for loan default risk assessment,” Economic Computation And Economic Cybernetics Studies And Research, vol. 53, no. 2, pp. 149–165, 2019.
6. Q. Yang and C. Wang, “A study on forecast of global stock indices based on deep LSTM neural network,” Statistical Research, vol. 36, no. 6, pp. 65–77, 2019.
7. K.-S. Moon and H. Kim, “Performance of deep learning in prediction of stock market volatility,” Economic Computation And Economic Cybernetics Studies And Research, vol. 53, no. 2, pp. 77–92, 2019.
8. J. Li, S. Pan, L. Huang, and X. Zhu, “A machine learning based method for customer behavior prediction,” Tehnicki Vjesnik-Technical Gazette, vol. 26, no. 6, pp. 1670–1676, 2019.
9. Aparna Nayak, M. M. Manohara Pai and Radhika M. Pai, “Prediction Models for Indian Stock Market”, (IMCIP) (2016).
10. Vijh, Mehar & Chandola, Deeksha & Tikkiwal, Vinay & Kumar, Arun. (2020). Stock Closing Price Prediction using Machine Learning Techniques. Procedia Computer Science. 167. 599-606. 10.1016/j.procs.2020.03,326
11. Hyeong Kyu Choi, “Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model”, arXiv:1808.01560v5 [cs.CE],(2018).
12. Pushpendu Ghosha, Ariel Neufeldb, Jajati Keshari Sahoo, “Forecasting directional movements of stock prices for intraday trading using LSTM and random forests”, (2021).
13. Mehtabhorn Obthonga, Nongnuch Tantisantiwong, Watthanasak Jeamwatthanacha c and Gary Willsd, “A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques”, (2020).
14. Wenjie Lu, 1, 2 Jiazheng Li, 3 Yifan Li, 3 Aijun Sun, and Jingyang Wang, “A CNN-LSTM-Based Model to Forecast Stock Prices”, (2020).
15. Nusrat Rouf, Majid Bashir Malik, Tasleem Arif, Sparsh Sharma, Saurabh Singh, Satyabrata Aich , and Hee-Cheol Kim, “Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions”, (2021).
16. Jingyi Shen and M. Omair Shafq, “Short-term stock market price trend prediction using a comprehensive deep learning system”, (2020).
17. D. Wei, "Prediction of Stock Price Based on LSTM Neural Network," International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), (2019).
18. Chen, Sheng & He, Hongxiang. (2018). Stock Prediction Using Convolutional Neural Network. IOP Conference Series: Materials Science and Engineering. 435. 012026. 10.1088/1757-899X/435/1/012026.
19. Wu, J.MT., Li, Z., Herencsar, N. et al. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. Multimedia Systems (2021). https://doi.org/10.1007/s00530-021-00758-w.
20. Ji, X., Wang, J. and Yan, Z. (2021), "A stock price prediction method based on deep learning technology", International Journal of Crowd Science, Vol. 5 No. 1, pp. 55-72. https://doi.org/10.1108/IJCS-05-2020-0012.
21. Vanukuru, Kranthi. (2018). Stock Market Prediction Using Machine Learning. 10.13140/RG.2.2.12300.7748.
22. A M Pranav, Sujooda S, Jerin Babu, Amal Chandran, Anoop S, 2021, StockClue: Stock Prediction using Machine Learning, International Journal Of Engineering Research & Technology (IJERT) NCREIS – 2021 (Volume 09 – Issue 13).