

# Optimizing the Duration for Feature Selection in LSTM Based Stock Price Prediction

**Author Details: Krishiv Shah**

Aditya Birla World Academy, Mumbai, India, krishiv.shah2004@gmail.com

## Abstract:

*The increase in availability of data and sophisticated machine learning methods, forecasting stock price has become a major attraction for both data scientists and traders alike. LSTM models are one of the most advanced in terms of determining patterns undetectable to the human minds and are utilized in stock price prediction for the same advantage. This paper evaluates the impact of number of training data points on the intraday returns forecasted using LSTM. Both the returns and the volatility is considered and the results are verified over a large duration and comparisons are made between the sizes of different training spans. High sharpe ratios ( $>2$ ) were obtained with multiple partition sizes with improved mean intraday returns. The partition size of 50 was found to be the most appropriate for stock price forecasting.*

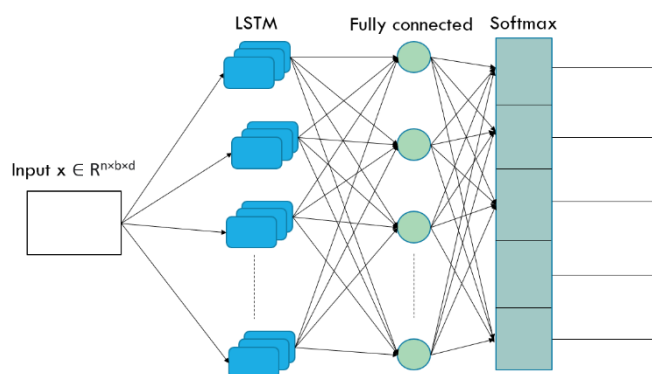
**Keywords:** LSTM, training span, Sharpe Ratio, Mean returns

## 1. Introduction

Many countries saw an exceptional rise in number of trading accounts in the covid era and it's a testimony of the growing interest of the masses in stock trading. With the increase in volatility of the markets, stock price prediction has gained traction in recent years with the further development of deep learning and data availability. This data availability has led to the quantification of several important factors which affect stock prices and enabled conditional formulization of their impact on stock prices. This prediction in stock prices is utilized in gauging both the future trends of a stock and the actual price. Initially, statistical models such as autoregressive integrated moving average (ARIMA) models were developed to forecast stock price; these were a product of intuition. However, they were ineffective in stock markets due to the nonlinear nature of growth/ decline. It is only with the growth of computational intelligence in recent years that non-linear models empowered by artificial neural networks (ANNs), fuzzy-neural systems, genetic algorithms, evolutionary and particle swarm methods have been employed by many researchers for stock market forecasting.

LSTM models are subsidiaries of Recurrent Neural Networks, most useful for sequencing operations such as time-series models, the primary object of this paper. Comprising 4 units, the cell, the input, the forget and output, the neural network is self-sufficient in selection of data to be considered and the data to be ignored for every instance. Additionally, it uses this mechanism to determine how the output of step  $n$  will be utilized in step  $n+1$  and in subsequent outputs. This makes it highly streamlined in processing newer data on a constant basis, and using the output from that data as influences on further input. Neural Networks are especially efficient in determining occurrences or patterns which occur after irregular intervals, essentially 'lagging' [1,2]. These patterns are often completely indiscernible to human intelligence. It is also this attribute of LSTM models which gives them an edge over hidden Markov Models and classic RNNs, avoiding vanishing gradient while backpropagation [3,4].

It is the aforementioned properties of LSTM models which makes them extremely suitable for applications in time-based predictions, particularly assets like stocks. The basic premise stems from the fact that all stock prices are affected in part by a variety of identifiable variables, which by their very nature repeat simultaneously at irregular intervals of time, having a similar effect on the price, the magnitude of their effect varying with their own magnitude. This clearly indicates the importance of historical data for the identification of variables and training the weights in the neural network, which is available in plenty when it comes to stocks. Thus, stocks are prime subjects of LSTM predictions, the quality of which is rivaled only by Gated Recurrent Unit Models [5,6,7]. Thus, the intelligent variation in data considered and dynamic addition to this data make LSTM models perfectly suited to stock price predictions.



A predictive model is based on a multitude of input variables. Predicting a phenomenon with a high number of input variables can not only increase the computation costs but also induce a certain degree of randomness. Therefore, in order to reduce the number of inputs and choose variables with higher influence, the process of feature selection is carried out [9,10]. In stock prediction, there is abundant ambiguity in determining the ideal duration of feature selection. Earlier researchers heavily relied on utilizing the partitions of one year (around 250 days) in the training spans to identify variables with higher weightage. This strategy could be useful for seasonal forecasting or industries with long business cycles [11]. At the same time, it could be detrimental for industries with dynamic business scenarios and fluctuating public demand. Therefore, it is essential to identify an optimized size of periods in which the training and feature selection spans can be divided. In this work, an attempt has been made to optimize the number of training days of the LSTM model in order to enhance the intraday returns obtained by trading in 47 stocks of the Nifty 50 group. This includes out of the 'n' chosen stocks, predicting n/2 stocks which would return a profit in the intraday transaction and the remaining n/2 stocks which would return a loss. The variation in number of different training data points is obtained by taking different sizes of partitions in the duration. These spans are evaluated on the basis of sharpe ratios as well as and the partition sizes with better returns and reduced volatility is deemed appropriate.

## 2. Materials and Methods

Stock prediction with LSTM can be divided into three major activities namely: Feature generation, training, and testing. The initial duration is utilized for selecting specific features and important variables. In order to determine the most suitable partition size, we have Figure 1 illustrates the division of the training span into partitions of 50 days and hence each year will have around 4-5 partitions, assuming 240 to 250 trading days. Similarly, as mentioned earlier, the accuracy of prediction is evaluated for 5 sizes of partitions and hence the number of partitions in one year will be different as the size of these partitions varies.

The specifications of the training spans are mentioned in table 1 assuming 245 trading days in a calendar year.

Table 1 Specifications of the training spans

Training Span	Size of partitions $\Delta t$	Number of training data points per annum
T1	20	207
T2	50	147
T3	100	47
T4	200	45
T5	240	5

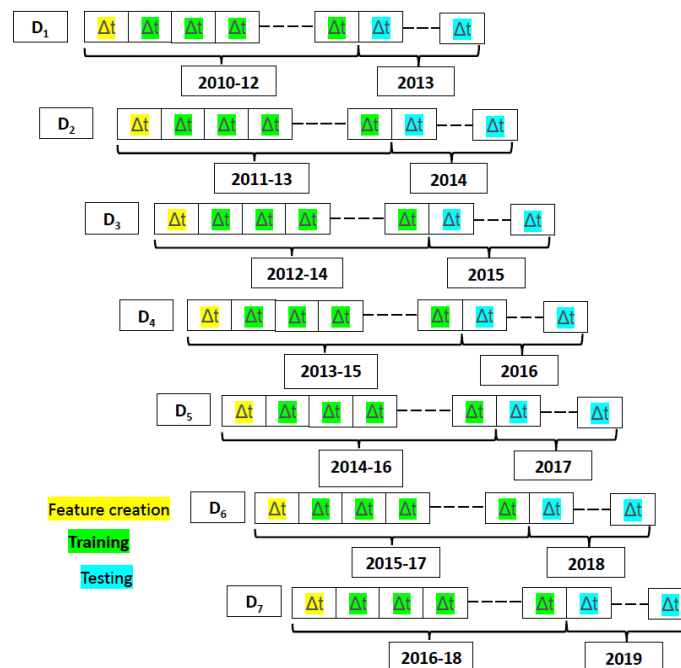


Figure 2 Sequence of feature creation, training and testing periods

Based on the prediction, the sharpe ratios and mean returns were calculated for different years for specific sizes of the partitions in training spans. A linear regression trendline is added to each plot which facilitates the visualization of the movement of the quantities.

The quantity sharpe ratio is defined as:

$$S_a = \frac{E(R_a - R_b)}{\sigma_a}$$

Where,

$S_a$  = Sharpe ratio

$E$  = expected value

$R_a$  = Asset return

$R_b$  = risk free return

$\sigma_a$  = standard deviation of asset access return

Intraday returns of for a stock can be expressed as:

$$R = \frac{C}{O} - 1$$

Here,

$R$  is the intraday returns obtained,

$C$  and  $O$  are the closing and opening values of a stock respectively

### 3. Results

It was observed that for most sizes of the partitions the values of sharpe ratios and means increased as the prediction progressed from 2013 to 2018. The partition size of 50 days was found to be the most appropriate when we compared the results of sharpe ratios and mean average returns. The change in sharpe ratios with time for the partition size 50 is compared with other sizes in figures 2,3,4, and 5.

From figure 2, it can be observed that, with a partition size 20, there is a steep decline in the sharpe ratio, for the initial 2 years and grows marginally for the next year followed by a steep ascend in the next year. This illustrates the volatility of prediction with the partition size of 20 and hence it was considered inappropriate for forecasting although the linear regression trendline depicts the overall increase in the sharpe ratio in the entire duration.

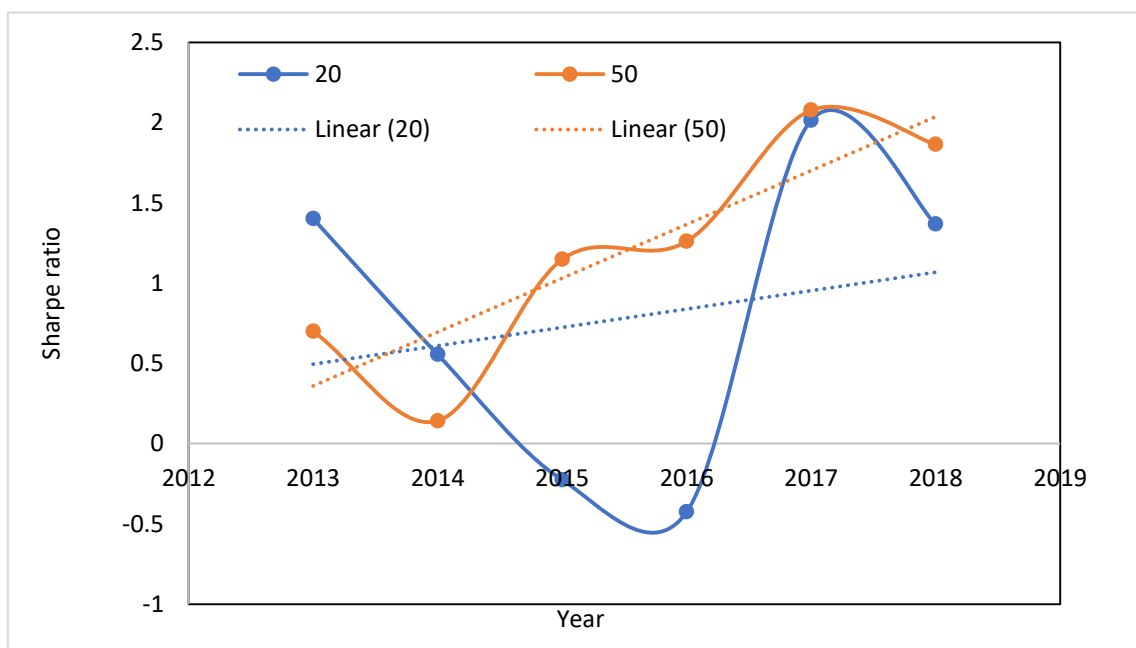


Figure 3 A comparison of sharpe ratios for partition sizes 50 and 20

When we analyze the sharpe ratios obtained with a partition size 100 (with a plot depicted in figure 3), we can notice that this partition size is the closest to the results obtained with 50, however, with 100, the results are slightly lower. This point can also be validated by observing the proximity in the two trendlines.

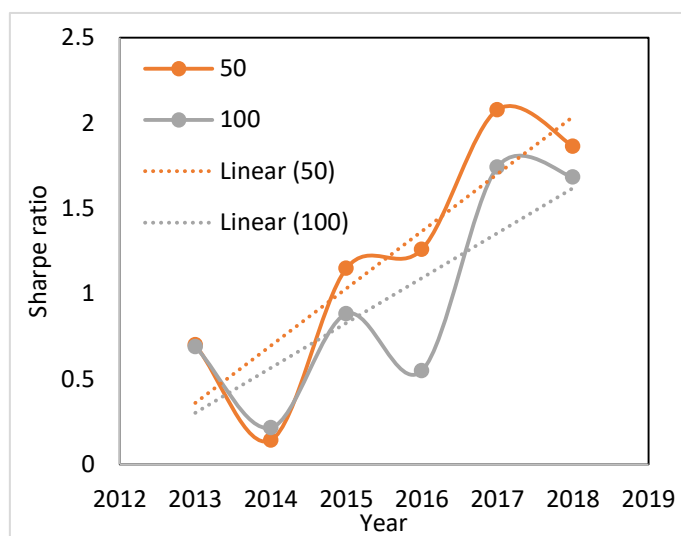


Figure 4 A comparison of sharpe ratios for partition sizes 50 and 100

The comparison between the partition sizes 50 and 200 in terms of sharpe ratio is demonstrated in figure 4. The sharpe ratios obtained with a partition size of 200 give an overall decline between 2013 and 2018, as predicted by the trendline.

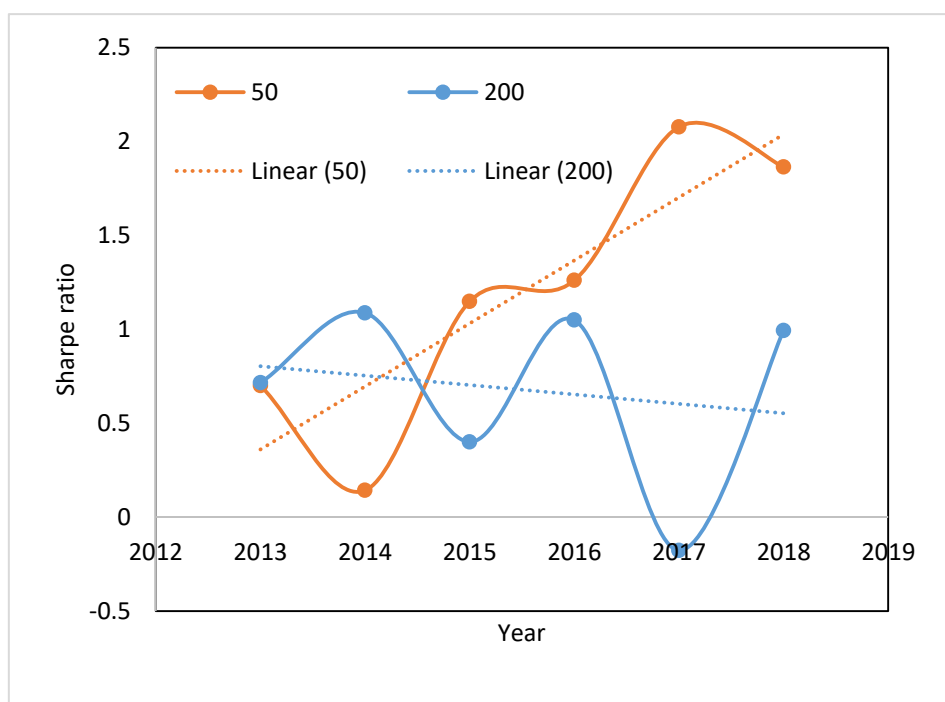


Figure 5 A comparison of sharpe ratios for partition sizes 50 and 200

A very volatile change in sharpe ratios was observed with the partition size 240, as shown in figure 5 and hence it was found unsuitable for the prediction of stock prices.

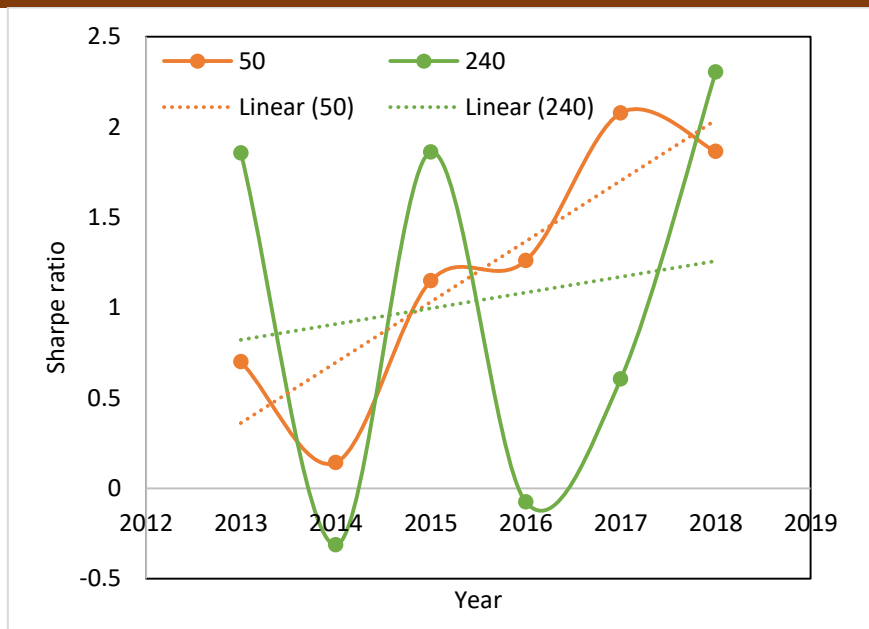


Figure 6 A comparison of sharpe ratios for partition sizes 50 and 240

A similar analysis was done for different partition sizes with the mean returns as well. It was observed that partition 50 was found to be suitable with this criterion as well. The comparisons are illustrated in figures 6,7,8 and 9.

As noticed in the case of the sharpe ratio, the amount of volatility was high in mean returns as well, with a partition size of 20 (figure 6).

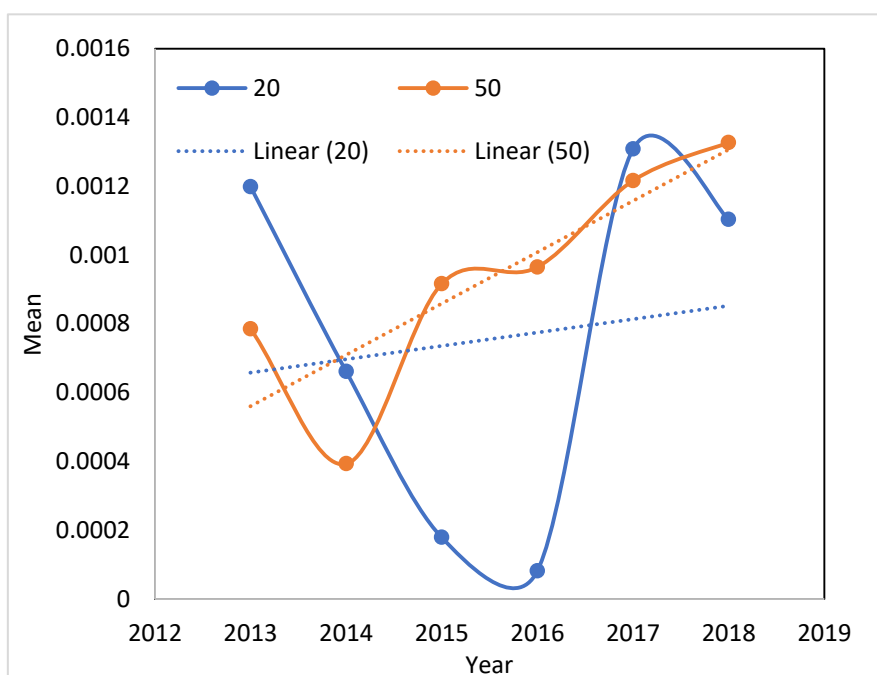


Figure 7 A comparison of mean returns for partition sizes 50 and 20

The similarity was found in the case of the partition size of 100 as well (figure 7), where the mean returns for the partition size 50 and 100 are close with the one with 50 being slightly higher.

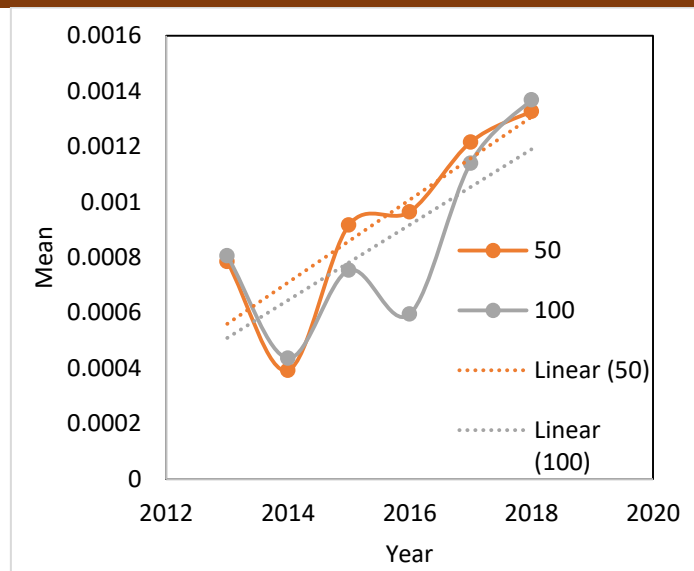


Figure 8 A comparison of mean returns for partition sizes 50 and 100

Although the mean returns with the partition size 200 rise steeply in the initial few years, the drop after the year 2015 is sharp as well (figure 8), and therefore, this partition size was not considered suitable for prediction.

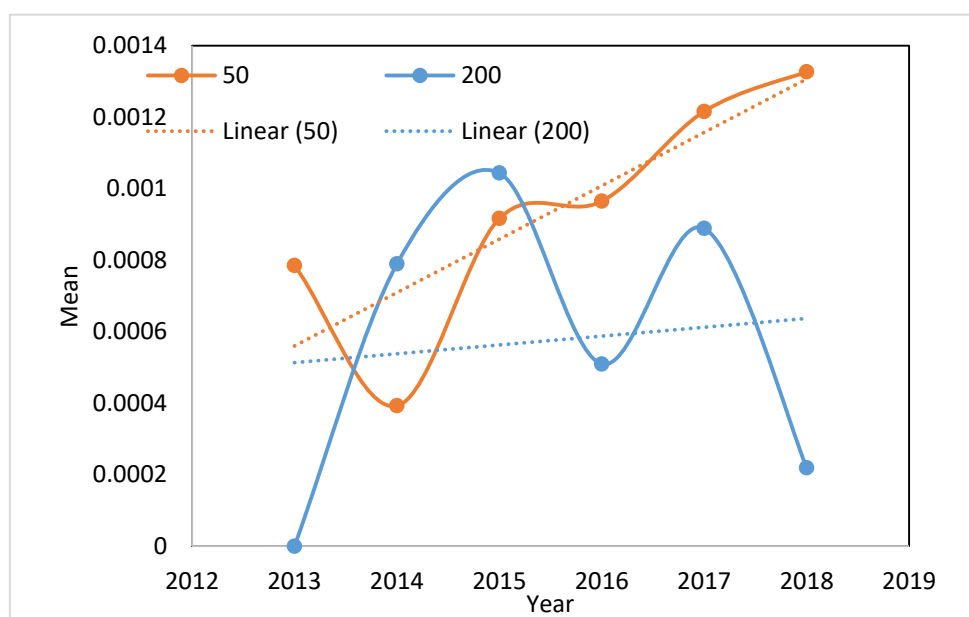


Figure 9 A comparison of mean returns for partition sizes 50 and 200

The volatility with the partition size 240 was also obtained the mean returns values (figure 9).

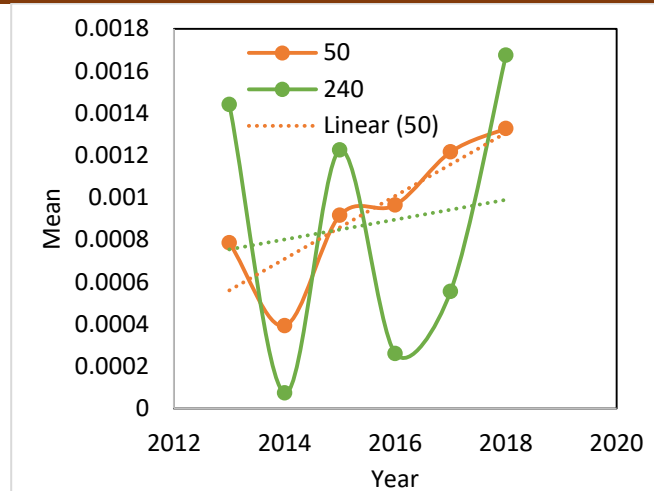


Figure 10 A comparison of mean returns for partition sizes 50 and 240

## 5. Conclusions

Five differently partitioned training spans of an LSTM based stock price prediction model were evaluated for their sharpe ratios and mean returns. The sizes of these partitions varied from 20 to 240 days. The training span with partitions of 50 days was found to be most suitable with better mean returns and sharpe ratio as compared to other durations in the consideration. Partition size of 100 gave the closest results to that of 50. The amount of volatility obtained in the results with other partition sizes was much higher and hence 50 was chosen to be the appropriate size.

**Funding:** This research received no external funding

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- i. Rybalkin, V., Sudarshan, C., Weis, C., Lappas, J., Wehn, N. and Cheng, L., 2020. Efficient Hardware Architectures for 1D-and MD-LSTM Networks. *Journal of Signal Processing Systems*, 92(11), pp.1219-1245.
- ii. Sharma, A., Tiwari, P., Gupta, A. and Garg, P., 2021. Use of LSTM and ARIMAX Algorithms to Analyze Impact of Sentiment Analysis in Stock Market Prediction. In *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020* (pp. 377-394). Springer Singapore.
- iii. Siami-Namini, S. and Namin, A.S., 2018. Forecasting economics and financial time series: ARIMA vs. LSTM. *arXiv preprint arXiv:1803.06386*.
- iv. Qiu, J., Wang, B. and Zhou, C., 2020. Forecasting stock prices with long-short term memory neural network based on attention mechanism. *PloS one*, 15(1), p.e0227222.
- v. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259, 689–702.
- vi. Krauss, C., Do, X.A. and Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, 259(2), pp.689-702.
- vii. Fischer, T. and Krauss, C., 2018. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), pp.654-669.

- viii. *Sumon, S.A., Shahria, T., Goni, R., Hasan, N., Almarufuzzaman, A.M. and Rahman, R.M., 2019, April. Violent Crowd Flow Detection Using Deep Learning. In ACHIDS (1) (pp. 613-625).*
- ix. *Cheng M, Cai K, Li M. Rwf-2000: An open large scale video database for violence detection. In 2020 25th International Conference on Pattern Recognition (ICPR) 2021 Jan 10 (pp. 4183-4190). IEEE.*
- x. *Das, S. and Mishra, S., 2019. Advanced deep learning framework for stock value prediction. International Journal of Innovative Technology and Exploring Engineering, 8(10), pp.2358-2367.*
- xi. *Ghosh, P., Neufeld, A. and Sahoo, J.K., 2021. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests. Finance Research Letters, p.102280.*