(IJTBM) 2023, Vol. No. 13, Issue No. I, Jan-Mar **e-ISSN: 2231-6868 p-ISSN: 2454-468X**

Developing an Integrated Smart Model to Enhance the Efficacy of Stock Market Prediction by Leveraging XGBoost and Long Short-Term Memory Networks

Arnav Goenka

Vellore Institute of Technology, Vellore, Tamil Nadu, India

DOI:10.37648/ijtbm.v13i01.010

[1]*Received: 18 December 2022; Accepted: 17 January 2023 ; Published: 11 March 2023*

ABSTRACT

A well-known economic tactic, the stock exchange has emerged as a crucial testing ground for the rapidly developing science of machine learning (ML). Stock prices can be predicted by using machine learning (ML) to analyse several aspects of the behaviour of the stock market. Given that stock prices are dynamic and influenced by real-time events, they cannot be predicted. However, deep learning algorithms can easily handle intricate data given in different patterns of stock prices.

The objective of this research was to predict closing stock prices for 25 companies listed on the Indonesia Stock Exchange (IDX). We employed two machine learning methods namely; Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) which have proven their worth in terms of predictive accuracy over a wide range of datasets. In addition, we developed a trading strategy using two thresholds that signal when it's best to buy or sell.

The outcomes of this experiment are very promising indeed. XGBoost algorithm achieved an impressive success rate with 99% prediction accuracy – a great performance for any model! This not only validates our methodology but also highlights the potential applicability of our study within real-world trading environment where accurate predictions are crucial.

INTRODUCTION

The stock market is one of the most liquid form of investments but at times regarded as difficult and unpredictable [1] . The complex and constantly changing data, graphs, candlesticks, indicators on the stock exchange can be intimidating to individuals who are not involved in the financial industry. This dynamic nature of stock market poses a barrier for traders seeking to devise lucrative trading techniques. Stock market data is one of the most challenging time series forecasting problems because it is volatile [2], [3]. Machine learning (ML) can now make accurate predictions by using the huge amounts of data generated by the stock market. By giving traders advice on timings of a trade like entry, exit and stoploss, they can ultimately maximise profits [4],[5] and [6]. A well-known theory in financial economics called the Efficient Market Hypothesis (EMH) states that asset prices represent all available information. Outperforming the market is a rare achievement under EMH, according to multiple research [1], [9], and [10]. There are three categories for this hypothesis: weak, semi-strong, and strong. The weak form implies that stock prices already take into account all past trade data. Price fluctuations are a result of the semi-strong form's ability to adjust to new information. Insider or private information can be incorporated into the strong form to create highly customised models for traders or businesses [10], [11]. Even though EMH is widely accepted, its findings are sometimes controversial [11]. Warren Buffet is a prominent counterexample to this concept, having regularly outperformed the stock market over an extended period of time. Engaging with the theoretical foundations of our study requires an understanding of the EMH's practical implications. The growth of automated strategies for trading stocks is linked to the rise of artificial intelligence. Traders can use AI to predict market trends, due to its capabilities to handle huge datasets. This theory goes hand in hand with

[1] *How to cite the article:* Goenka A. (March, 2023); Developing an Integrated Smart Model to Enhance the Efficacy of Stock Market Prediction by Leveraging XGBoost and Long Short-Term Memory Networks; *International Journal of Transformations in Business Management,* Vol 13, Issue 1, 110-117, DOI: http://doi.org/10.37648/ijtbm.v13i01.010

Warren Buffet's method of predicting where markets are headed by considering as much data as possible.[3], [7], [12].

Python was our programming language of choice for this work because of its flexibility with data and the availability of pre-built models [3], [7], [12]. We specifically used the Long Short-Term Memory (LSTM) model, whose memory component makes it excellent in time series forecasting. To assess LSTM's efficiency in stock price forecasting, we also contrasted it with another machine learning method, the Extreme Gradient Boosting (XGBoost) algorithm.

To ensure that our research was relevant, we studied the historical stock market data of 25 companies enlisted in the Indonesia Stock Exchange (IDX). These machine learning systems were trained using past data to predict future stock prices. This focus on real-world data guarantees the relevance and dependability of our conclusions, boosting our approach's efficiency.

MATERIALS AND ALGORITHMS

This section describes the dataset used, its importance, and the machine learning techniques that were used: Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM).

Stock Exchange Dataset Used

Our goal is to predict the closing prices of stocks of 25 companies that are listed on the Indonesia Stock Exchange (IDX). The dataset, which is a special compilation of daily adjusted data for the years 2000–2019, was acquired using the public dataset site Kaggle [13]. This dataset is a novel and important resource in the field of stock market prediction and serves as the foundation for our study. To train our models, we made use of historical data from the preceding n days.

Fig. 1. Plotting stock from each company in single days

Various market indicators, such as minimum and maximum prices, opening and closing prices, and trading volumes, are used to predict stock prices [8], [13]. Figure 1 illustrates an example of foreign stock data over days.

XGBoost: The Meta Algorithm for Stock Market Prediction

Extreme Gradient Boosting (XGBoost) is one of the most popular machine learning techniques [14], [15]. It has consistently demonstrated impressive performance across various datasets [14], [16]. What it does is that it transforms weak classifiers into more robust ones through iterative training. This process is repeated to build a strong model for stock market prediction. "boosting" refers to enhancing the performance of tree-based algorithms by maximizing computational resources. Since its introduction in 2014 [15], XGBoost has been widely adopted for its outstanding performance in numerous experiments.

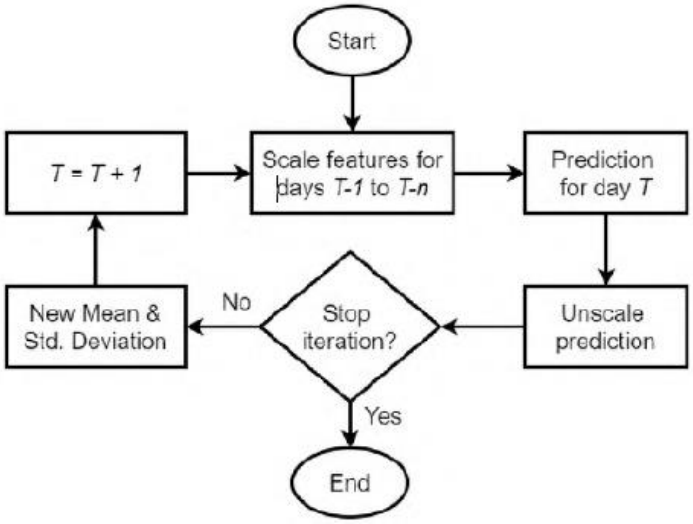


Fig. 2. Logic implementation by recursive forecasting

The developer of XGBoost, Tianqi Chen, claimed notable gains in performance by ensemble-ing a group of Classification And Regression Trees (CART) using a gradient descent technique [16]. By integrating the prediction capacity of several decision trees, CARTs perform better than single trees. The trees essentially serve as a jury, casting their aggregate votes for the most likely result. To minimise loss, XGBoost iteratively adds new models and fixes the mistakes in older models [16].

An artificial model that originates from a single root node and branches out into several possible outcomes is called a decision tree. Every branch shows the result of a test conducted on a particular feature, every non-leaf node a test conducted on that feature, and every leaf node a classification. Iteratively, the splitting process proceeds until a termination condition is satisfied. At each node, the ideal split is determined by minimising loss.

In boosting techniques, the additive training approach is used, where a weak classifier is added to the model in each phase. The loss function measures the predictive accuracy, while the regularization term controls overfitting. By iteratively applying weak classifiers and minimizing the loss function, the ensemble of weak classifiers (WCs) forms a strong classifier (SC), significantly improving predictive performance over random classifiers [16].

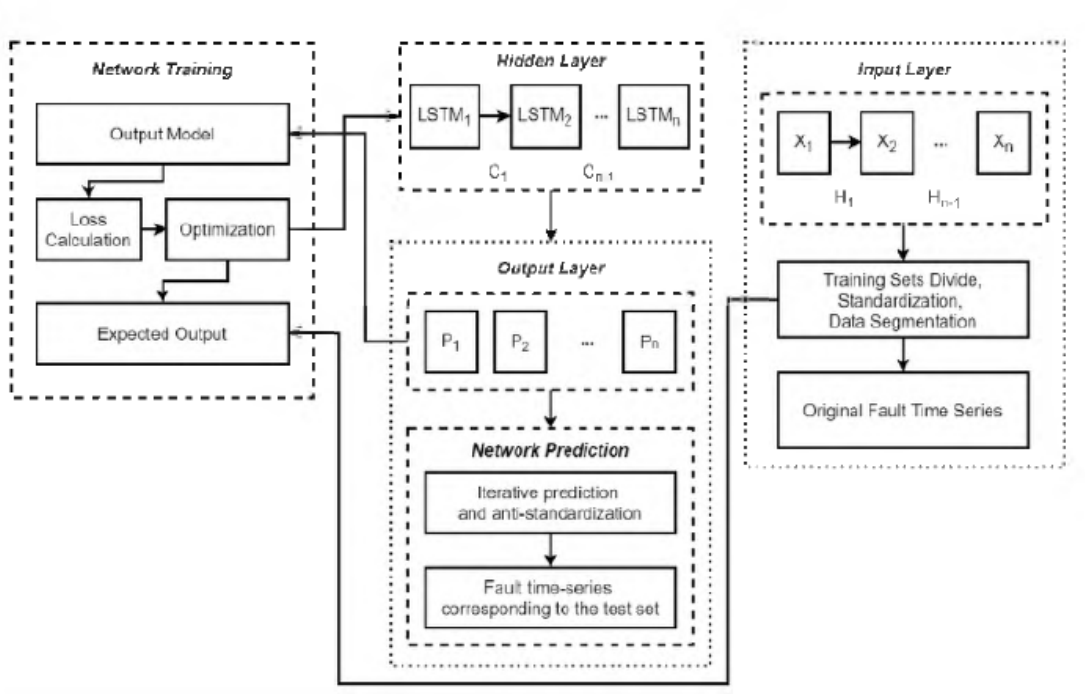


Fig. 3. General LSTM Architecture [17]

LSTM Architecture for Stock Market Modelling

The Long Short-Term Memory (LSTM) algorithm is an advanced type of Recurrent Neural Network (RNN) that has become very popular in many communities [18]-[20]. It is particularly effective at learning time series-based forecasting with long-term dependencies, and works well on large datasets. Researchers have made significant modifications to LSTM over the years to improve its capabilities.

LSTM architecture differs from normal RNN structures which look like chains looping back into previous layers by having four interacting network layers. One of its most important feature is the "forget" gate mechanism which allows it to backpropagate errors through many layers. This means that it can remember things for a long time, reducing error rates along the way by learning from past steps. Figure 3 illustrates the general LSTM architecture.

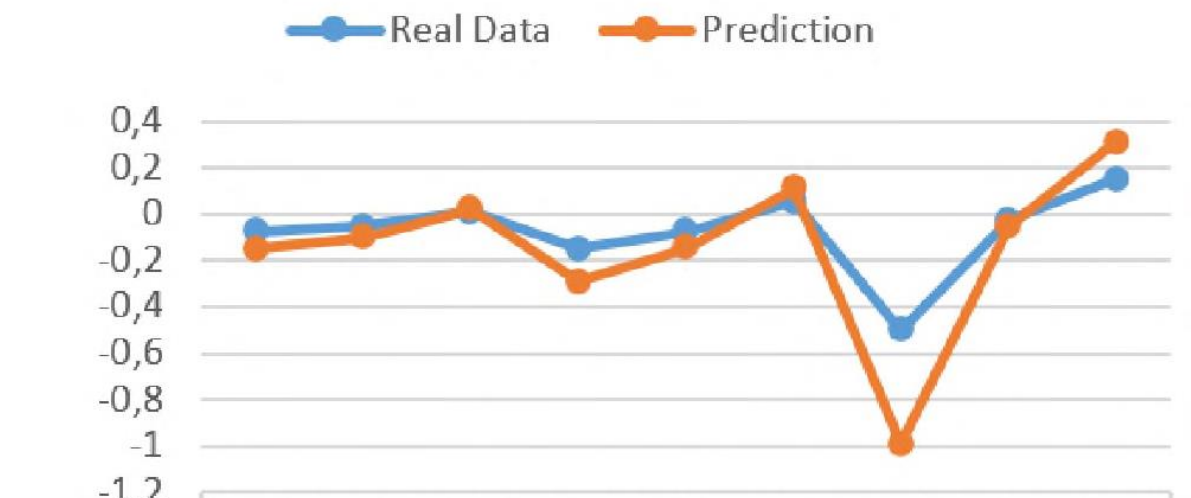
EXPERIMENTAL RESULTS

Before applying XGBoost and LSTM to our data, we divided the dataset into training, validation, and testing subsets. Both algorithms' training data used for parameter tuning comprised 60% of the total instances. The validation data used for model tuning accounted for 20%, and the testing data used to evaluate the final models also accounted for 20%.

Notwithstanding the widely recognized difficulties of deep learning (DL) because of its expansive architectures and data-driven nature, which usually makes it take up much computation time and expensive in terms of resources, this research was successful to address these problems by employing both XGBoost and LSTM.

Prediction using XGBoost

In our experiment with XGBoost, the strategy was as follows: we built the learning model using the training data, tuned the hyperparameters using the validation data, and evaluated the final model using the testing data. This process ensured that the model was optimized and could deliver accurate predictions.



	2016-01	2016-05	2016-09	2017-01	2017-05	2017-09	2018-01	2018-05	2018-09
Prediction	-0,07	-0,05	0,012	-0,14	-0,07	0,055	-0,49	-0,03	0,159
Real Data	-0,08	-0,05	0,01	-0,15	-0,07	0,055	-0,5	-0,03	0,15

Fig. 4. Plot result shows the predictions using XGBoost

The applied features for our models included open prices, maximum prices, minimum prices, closing prices, and trading volumes. The data for n days was appropriately split into training, validation, and testing sets. Figure 4 shows the results of the XGBoost stock market prediction, yielding a 99% accuracy. The vertical axis represents

the monthly Return on Investment (ROI), while the horizontal axis represents the time variable in YYYY/MM format.

In contrast, Figure 5 displays the error for the XGBoost algorithm, measured using the Root Mean Square Error (RMSE) metric. The RMSE is presented for iterations over 60 days and 90 days.

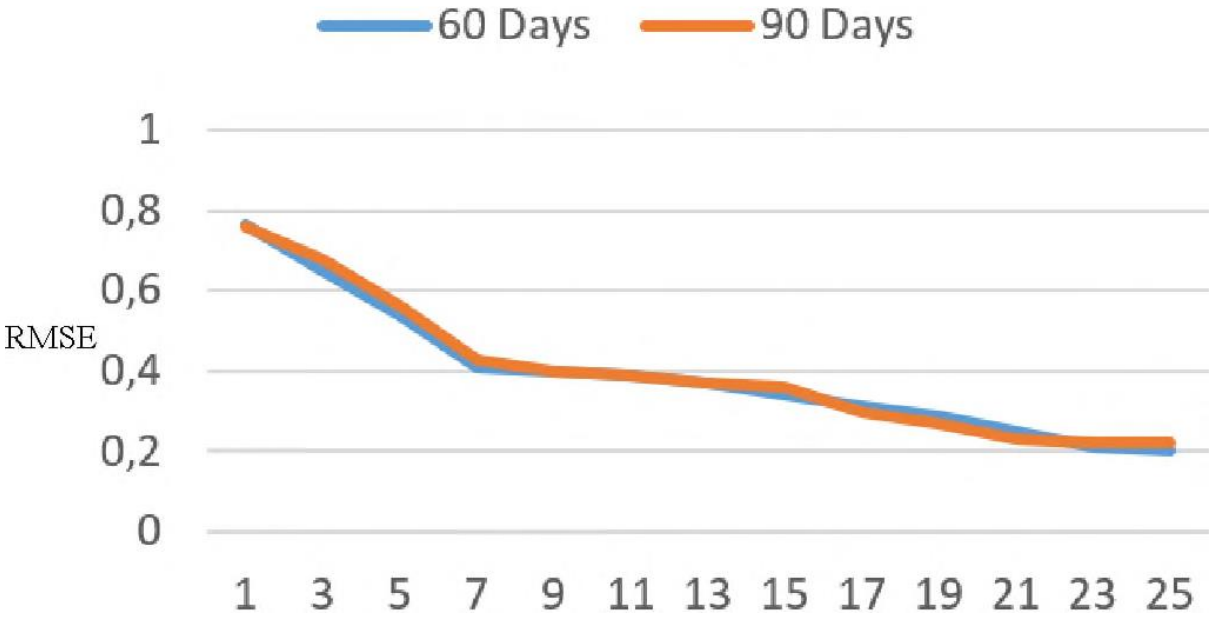


Fig. 5. Error chart in XGBoost using RMSE evaluation metric

Prediction using LSTM

LSTM Based Prediction has emerged as a deep learning-based methodology specifically designed for overcoming the issues of vanishing gradients in long-term dependencies. It is different from other methods due to its architecture design with three gates: input gate (also called update gate), forget gate and output gate. The purpose of these gates is to decide what information is store or discard in memory cell at each time step [17][21]. Therefore, it can remember useful things even over very long sequences which makes this algorithm very powerful when used for stock market forecasting.

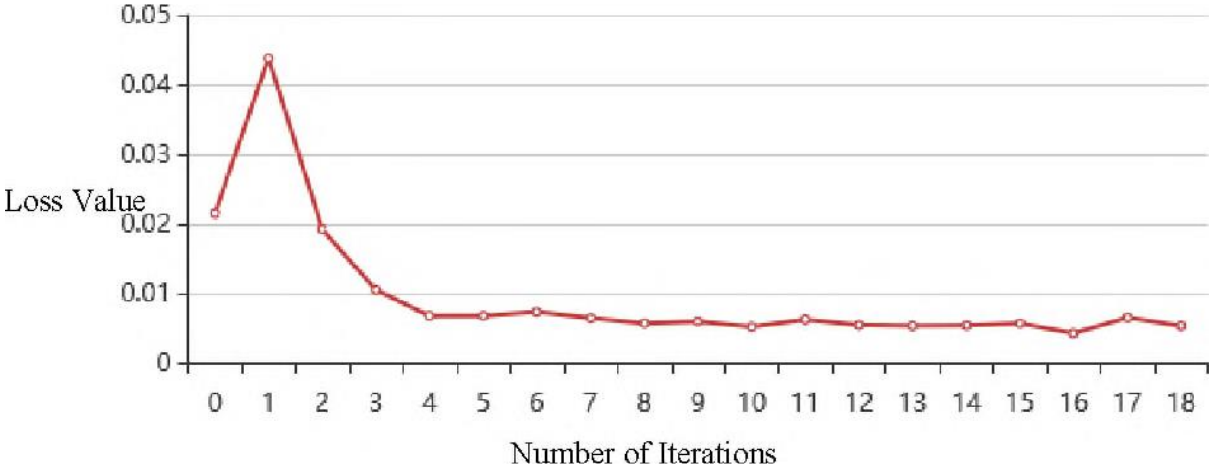


Fig. 6. Loss graph using LSTM

The output gate, the third gate in the LSTM architecture, determines how much information from the memory cell is used as output to activate the next layer. To prevent overfitting, a dropout layer is placed between two LSTM layers.

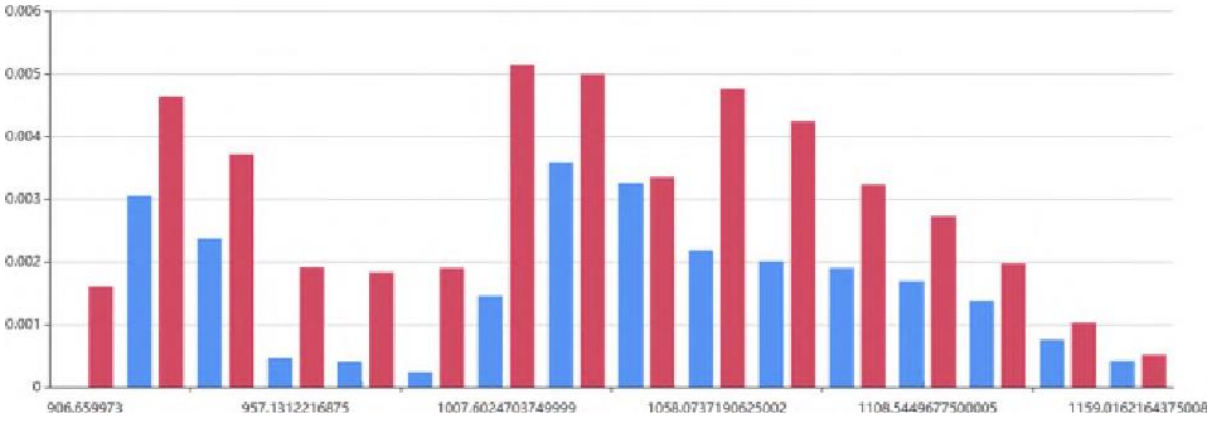


Fig. 7. Histogram chart of real data and predicted data

Figure 6 displays a loss graph showing a minimal loss of 0.005% by the last epoch. For LSTM, we conducted 10 epochs/learning iterations, and the improvement can be observed in Figure 6. In Figure 7, a histogram chart compares real data with predicted data. Additionally, we simulated buying and selling based on our predicted histogram units.

TABLE I. Buy and Sell Simulation with LSTM Prediction

Date & Month (in 2017)	Action	Price (USD)	Investment Value	Balance
7 Sept	Buy	4,679.75	0%	5,323.899535000001
14 Sept	Buy	4,625.55	-1.15%	9,949.44946
20 Sept	Buy	4,657.91	0%	5,291.549375
25 Sept	Sell	4,604.85	-1.13%	9,896.399229999999
11 Oct	Buy	4,946.25	0%	4950.149229999999
16 Oct	Sell	4,960	0.27%	9,910.149229999999
24 Oct	Buy	4,852.7	0%	5,057.449339999999
1 Nov	Sell	5,127.5	5.66%	10,184.94934
13 Nov	Buy	5,128.8	0%	5,056.199339999999

TABLE II. Performance Comparison from Previous Work

Author	Year	Accuracy	
		XGBoost	LSTM
Dey et al. [22]	2016	88%	-
Vargas et al. [23]	2017	-	65.08%
Roondiwala et al. [18]	2017	-	99.92%
Chatzis [16]	2018	45%	-
Cai et al. [24]	2018	-	66%
Nobre and Neves [25]	2019	49.26%	-
Basak et al. [26]	2019	94.79%	-
This proposed experiment	2020	99%	99.995%

Table I illustrates the simulation of buy and sell units for one company using our predicted model throughout the year 2017. We fixed the transaction size at five units, with an initial investment of USD 10,000.

Table II compares the accuracy results from this study with those of previous works.

CONCLUSION

Developing a successful trading strategy is essential in the stock market, which operates as an investment instrument rather than a game of chance. Random decisions in buying and selling units are insufficient; a coherent profit strategy is crucial.

Recent advancements in Machine Learning (ML) have significantly benefited traders by shedding light on various aspects of the stock market.Many investigations have been made in the selection of best technical indicators, comprehension of market behavior and determination of effective strategies with the help of relevant market data. Out of all these, stock market prediction appears to be the most useful for traders.

In our investigation, we used two ML algorithms namely Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory (LSTM) to simulate stock exchange. All the algorithms performed remarkably well by showing high precision when forecasting prices of stocks.

REFERENCES

[1] E. Chong, C. Han, and F. C. Park, “Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies,” ExpertSyst. Appl., vol. 83, pp. 187-205, Oct. 2017.

[2] Y. S. Abu-Mostafa and A. F. Atiya, “Introduction to Financial Forecasting,” Appl. Intell., vol. 6, no. 3, pp. 205-213, Jul. 1996.

[3] E. L. de Faria, M. P. Albuquerque, J. L. Gonzalez, J. T. P. Cavalcante, and M. P. Albuquerque, “Predicting the Brazilian Stock Market through Neural Networks and Adaptive Exponential Smoothing Methods,” Expert Syst. Appl., vol. 36, no. 10, pp. 12506-12509, Dec. 2009.

[4] A. B. Gumelar et al., “Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks,” IEEE 7th Int. Conf. Serious Games Appl. Heal., pp. 18, Aug. 2019.

[5] D. P. Adi, A. B. Gumelar, and R. P. Arta Meisa, “Interlanguage of Automatic Speech Recognition,” in 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019, pp. 88-93.

[6] A. B. Gumelar, D. A. Lusia, A. Widodo, and R. Felani, “Using Neural Networks on Cloud Container’s Performance Comparison By R on Docker (ROCKER),” 2018 Int. Symp. Adv. Intell. Informatics, p. 5, 2018.

[7] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” J. Comput. Sci., vol. 2, no. 1, pp. 1-8, Mar. 2011.

[8] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, “Evaluating Multiple Classifiers for Stock Price Direction Prediction,” Expert Syst. Appl., vol. 42, no. 20, pp. 7046-7056, Nov. 2015.

[9] C.-H. Cheng, T.-L. Chen, and L.-Y. Wei, “A Hybrid Model based on Rough Sets Theory and Genetic Algorithms for Stock Price Forecasting,” Inf. Sci. (Ny)., vol. 180, no. 9, pp. 1610-1629, May 2010.

[10] A. Timmermann and C. W. J. Granger, “Efficient Market Hypothesis and Forecasting,” Int. J. Forecast., vol. 20, no. 1, pp. 15-27, Jan. 2004.

[11] B. G. Malkiel, “The Efficient Market Hypothesis and Its Critics,” J. Econ. Perspect., vol. 17, no. 1, pp. 59-82, Feb. 2003.

[12] R. Cervello-Royo, F. Guijarro, and K. Michniuk, “Stock Market Trading Rule based on Pattern Recognition and Technical Analysis: Forecasting the DJIA Index with Intraday Data,” Expert Syst. Appl., vol. 42, no. 14, pp. 5963-5975, Aug. 2015.

[13] A. Wibowo, “IDX Indonesia Stock Index Price,” 2019. [Online]. Available: https://www.kaggle.com/aufaawibowo/idx-indonesia- stock-price/data. [Accessed: 01-Jan-2020].

[14] P. Carmona, F. Climent, and A. Momparler, "Predicting Failure in the U.S. Banking Sector: An Extreme Gradient Boosting Approach," Int. Rev. Econ. Financ., vol. 61, pp. 304-323, May 2019.

[15] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme Gradient Boosting as a Method for Quantitative Structure-Activity Relationships," J. Chem. Inf. Model., vol. 56, no. 12, pp. 2353-2360, Dec. 2016.

[16] S. P. Chatzis, V. Siakoulis, A. Petropoulos, E. Stavroulakis, and N. Vlachogiannakis, "Forecasting Stock Market Crisis Events using Deep and Statistical Machine Learning Techniques," Expert Syst. Appl., vol. 112, pp. 353-371, Dec. 2018.

[17] L. Lv, W. Kong, J. Qi, and J. Zhang, "An Improved Long ShortTerm Memory Neural Network for Stock Forecast," MATEC Web Conf., vol. 232, p. 01024, Nov. 2018.

[18] M. Roondiwala, H. Patel, and S. Varma, "Predicting Stock Prices Using LSTM," Int. J. Sci. Res., vol. 6, no. 4, 2017.

[19] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock Price Prediction using LSTM, RNN and CNN-sliding Window Model," in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 1643-1647.

[20] D. Shah, W. Campbell, and F. H. Zulkernine, "A Comparative Study of LSTM and DNN for Stock Market Forecasting," in 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4148-4155.

[21] A. B. Gumelar, Eko Mulyanto Yuniarno, Wiwik Anggraeni, Indar Sugiarto, A. A. Kristanto, and M. H. Purnomo, "Kombinasi Fitur Multispektrum Hilbert dan Cochleagram untuk Identifikasi Emosi Wicara," J. Nas. Tek. Elektro dan Teknol. Inf. , vol. 9, no. 2, pp. 180-189, May 2020.

[22] S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting," PESITSouth Campus, 2016.

[23] M. R. Vargas, B. S. L. P. de Lima, and A. G. Evsukoff, "Deep Learning for Stock Market Prediction from Financial News Articles," in 2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), 2017, pp. 60-65.

[24] S. Cai, X. Feng, Z. Deng, Z. Ming, and Z. Shan, "Financial News Quantization and Stock Market Forecast Research Based on CNN and LSTM," 2018, pp. 366-375.

[25] J. Nobre and R. F. Neves, "Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to Trade in the Financial Markets," Expert Syst. Appl., vol. 125, pp. 181-194, Jul. 2019.

[26] S. Basak, S. Kar, S. Saha, L. Khaidem, and S. R. Dey, "Predicting the Direction of Stock Market Prices using Tree-based Classifiers," North Am. J. Econ. Financ., vol. 47, pp. 552-567, Jan. 2019.