JURNAL INFORMATIKA
PROXIES

Vol.9, No.1, July, 2025

Stock Prices Prediction Using Machine Learning

Selly Margaretha Sudiyandi[1], Robertus Setiawan Aji Nugroho[2]

[1] Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, sellymargaretha41@gmail.com
[2] Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, nugroho@unika.ac.id

Corresponding Author Email: nugroho@unika.ac.id

https://doi.org/10.24167/proxies.v9i1.13269

Received: 2025-02-19
Revised: 2025-10-22
Accepted: 2025-10-22
Available online:2025-10-23

Keywords:
stock prices, regression, lstm, cnn, lssvm

ABSTRACT

Prediction of stock price movements in the future will be an area that is widely researched. There is a hypothesis that it is considered impossible to predict stock prices, but it can also show that stock price forecasts can achieve a fairly high level of accuracy if properly formulated and modeled. This is because equity trading is one of most important investment activities. Modeling and forecasting future stock prices based on current financial information can be very helpful to investors. They want to know if inventories go up or down in the short or long term. In this research, the author wants to analyze the comparison of accuracy and train the dataset using linear regression, lasso regression, LSSVM, LSTM, and CNN, then the accuracy will be calculated from the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). This can be used by investors in predicting stock prices using a more accurate algorithm. The findings reveal that the CNN model has a substantially lower accuracy value, while LSTM also performs well on specific datasets. However, there is one difference between these two models: the LSTM training time is slower than the CNN model. This is because computations in CNNs may occur in parallel (the same filter is applied to numerous circumstances at the same time), but LSTMs need to be processed sequentially, because the next step depends on the prior time..

BACKGROUND

Prediction of stock price movements in the future will be an area that is widely studied. Although there is a hypothesis that it is believed to be impossible in predicting stock prices, there are also ways that show if formulated and modeled correctly, it will get a fairly high level of accuracy in stock price predictions. This is because stock trading is one of the main investment activities. Based on current financial information, modeling and predicting future stock prices is very useful for investors. They want to know whether a stock will go up or down in the short or long term.

One solution to overcome problems in stock price prediction for investors is to make machine learning by comparing the error value of one algorithm with another algorithm. The method used in predicting accuracy is calculating the value of Mean Absolute Error, Root Mean Squared Error, and Mean Absolute Percentage Error. Therefore, based on the calculated value, an algorithm that produces the lowest error value can be found.

The author wants to analyze the comparison of accuracy using Linear Regression algorithm, Lasso Regression algorithm, LSSVM algorithm, LSTM, and CNN to predict stock prices. The author wants to know which algorithm will get higher and more accurate accuracy. This **can be used by investors in predicting stock prices using a more accurate algorithm.**

LITERATURE STUDY

The type of data used in time series analysis is a way to break down the sequence of data points collected at a specific time. The difference between time series data and other data is how the variable changes over time, which shows the importance of time towards the adjustment of data points and the final result of the data. So the time series has two components, namely value and time. The goal is to observe or model existing data sets to enable accurate predictions of unknown future data values. The time series takes existing data and predicts the data's value.

Then there was some study that employed time series as the foundation. Mehtab and Sen [1]. This study examines the movement of stocks, particularly in the future, which is extremely difficult to forecast. The authors provide a methodology for forecasting stock prices that is extremely dependable and accurate, based on a mix of statistics, deep-learning models, and machine-learning. The **authors utilized day by day stock price data from a renowned corporation that is recorded on India's NSE, which is collected at five-minute intervals. The granular data's three daily slots are used to aggregate the data, which is then used to train and develop forecasting models. The authors contend that the agglomerative model construction methodology, which** integrates statistical, machine learning, and deep learning methodologies, may successfully learn from erratic and turbulent movement arrangement stock price data. CART technique generates decision trees that are totally binary in order to assure that each node has exactly two branches. The method separates the records in the training dataset into groups of records that have comparable values for the required characteristics. The trees are constructed by conducting a comprehensive search on each node for all accessible variables and potential splitting values, then selecting the best split based on some goodness of split criterion. Based on the measure of the proportion of RMSE to an average of the actual values of the projected variable, CNN models have been demonstrated to be even more precise than machine-learning models and LSTM models.

Pavlyshenko [2]. The author examines how machine learning models are utilized in sales predictive analytics in this paper. The fundamental purpose of this research is to look into the key methodologies and instances of utilizing machine-learning towards forecast sales. It has been discussed how machine learning generalization would impact things. Despite the fact that a limited amount of data is accessible for a certain sales period, like when a new shop or item is launched, such impact may be used to generate sales projections**.** Stacking single models to build regression ensembles has been researched. For our research, the author examined historical shop sales data from the Rossmann Store Sales. The findings show that by utilizing stacking methods, the author may enhance the effectiveness of statistical models for sales time series prediction.

And last, Namini et.al [3]. The purpose of this study is to determine if and how newly found deep-learning-based time series prediction models may be used., such as LSTM outperforming previously established ones. According to actual studies that have been undertaken and reported, LSTM and other deep learning-based algorithms outperform further commonly-based algorithms such as the ARIMA model. More precisely, LSTM obtained an average error rate reduction that was among 84 and 87 percent decreased than ARIMA, verifying LSTM excellence over ARIMA.

Following the research that utilized time series as the basis approach, there are several studies that use machine learning as the base method. Milosevic [4]. The author of this research provides a machine-learning aided approach to predicting the long-term future price of a share. In 76.5% of situations, the technique can predict whether the value of a specific firm will rise by 10% or not over the course of a year. The author trained the model in the Weka toolbox using a scope of classification machine-learning methods. First, in a 10-fold cross validation for all of these techniques, the author used all of the indicators and previous prices as features. The author then performed manual feature selection by removing attributes and determining if algorithm performance improved or worsened. The dataset was obtained using the Bloomberg terminal.

Sen [5], The author utilizes statistical, deep-learning, and machine-learning approaches to build a trustworthy and precise framework for stock price **forecast. The NSE India offers daily data on stock prices at five-minute intervals, which the author** utilizes to construct a stock price forecasting framework. The author claims that this system may be utilized for short-term stock price forecasting since it incorporates a variety of deep-learning and machine-learning approaches to accurately assess stock price volatility. Eight classification and regression models have been developed using data from two NSE equities, Tata Steel and Hero Moto, including one that employs a deep learning-based technique. Extensive results on the efficiency of these eight classification models and eight regression models have been published. While ANN were discovered to have the highest degree of accuracy when employed as classification approaches, LSTM surpasses all regression models by a wide margin. The results have greatly improved our understanding of stock price forecasting. Classification-based procedures and other regression methodologies that were previously utilized to forecast price movement are no longer applicable.

Mussumeci and Coelho [6]. To anticipate dengue extent in 790 Brazilian cities every week, the author discusses and contrasts machine learning techniques incorporating feature selection, such as LASSO, RF, LSTM. To document the spatial aspect of disease transmission, the author employs multivariate time-series predictors as well as time-series from nearby cities. When the models were tested, the LSTM showed the lowest prediction errors when forecasting dengue outbreaks outside of the sample in cities of various sizes. The findings provided here supplement an exceedingly restricted corpus of work on the utilization of machine learning algorithms to predict disease outbreaks. The author demonstrated how deep-learning models, such as LSTM, may be used well in complicated prediction tasks. In this situation, the author wants to see a broader examination and application of machine-learning approaches. The LSTM model will be used in the Info dengue project to provide estimates of dengue prevalence in Brazilian cities.

Abdualgalil et al. [7]. To achieve this purpose, seven distinct models are used in this study: RF, XGboost, SVR, MLP network, CNN-LSTM Hybrid, LSTM, SARIMAX are some of the algorithms used. Models with the lowest amount of errors are then utilized to forecast a number of positive, number of negative, and TB occurrence cases in the lungs. Based on the testing results, the CNN-LSTM and MLP were chosen to anticipate bad pulmonary, clear pulmonary, and TB occurrence cases from 2020 until 2029 since they had a lowest prediction error compared to the other models. As a result of the forecasting, 117.861557 new pulmonary negative events were expected, 414.4704 new cases of tuberculosis and 153.029385 new pulmonary positive occurrences.

Mehtab et al. [8], By creating numerous deep-learning and machine-learning, the author of this paper proposes a hybrid modeling technique for predicting stock price. For our investigation, the author used NIFTY 50 index data from NSE India between December 29th, 2014 and July 31st, 2020. The author developed eight regression techniques using training data consisting of NIFTY 50 index records from December 29th, 2014 to December 28th, 2018. Using these regression models, the author anticipated the NIFTY 50 open values from December 31st, 2018 to July 31st, 2020. Building four deep-learning based on regression models utilizing LSTM networks and an unique walk-forward validation technique would then improve the predicted accuracy of our **forecasting system.** The hyperparameters of the LSTM models are tweaked using a grid-searching approach to ensure that validation losses normalize with escalating epochs and validation accuracy converges. The author employs four separate models, each with a unique architecture and input data format, to capitalize on the predictive ability of LSTM regression models to forecast future NIFTY 50 open values.

All of the regression models offer comprehensive results on a wide range of metrics. The most accurate model is the LSTM base univariate approach, which utilizes each week's historical data as input to anticipate the open price of the NIFTY 50 for the coming week.

Bastos [10]. The implementation of machine learning algorithms for the early detection of Bear Markets and other significant decreases in stock prices is the subject of this study. These algorithms will be employed (and refined), and it intends to construct **models that, in some way, comply with the principles proposed for this study - predict market drops in advance. The** classification process will run through as many iterations and splits of the data as the user selects, using a time-series cross-validation approach and out-of-sample data testing. The objective of this study is the usage of machine learning algorithms to identify Bear Markets and other large dips in stock prices early. These algorithms will be applied (and improved), and it is hoped that by incorporating economic factors, it would be possible to develop models that comply with the principles proposed for this study - forecast market drops in advance. To present the results, the method plot results() will be used, which generates graphics with the results for all models throughout time, from 1970 to 2019, with market decreases and accompanying delays mentioned. Additionally, the confusion matrix, accuracy, precision, recall, AUC, and AUTP findings will be calculated using method metricsResults().

Then there are several studies that employ least squares support vector machines (LSSVM). Ismail and Shabri (2014) [9]. This study compares it to other models from previous research, such as ANN, ARIMA, SVR, etc. There have been multiple types of forecasting models produced, and academics have relied on statistical methodologies to anticipate the future. This study examines the use of LSSVM models for Canadian Lynx forecasting. SVM is useful in time series forecasting because it can tackle nonlinear regression estimation problems. SVMs are a type of machine learning approach differentiated through the use of the kernel function and capacity control of the decision function. The suggested model's predictive capacity is compared to many other models such as ANN, ARIMA, and others. In this study, MAE and MSE, which also are commonly used to evaluate the results of time series prediction, were utilized as performance metrics.

Triyono et al. [11] examined the accuracy of up to 10 artificial intelligence algorithms in predicting stock price fluctuations. A time series technique is used to forecast stock prices, which is exceedingly difficult to execute without the support of computing technology. This study was carried out in stages, with the most recent being a distribution of training data. The first step will be carried out using 400 lines of training data, the second stage with 800 lines of training data, and the third stage with 1235 lines of training data. According to the findings of this study, the more historical data there is, the more accurately predicting outcomes. This is demonstrated by examining the reduced error value (MSE) after multiplying historical data. On historical data training of 1200, the LSSVM approach has the highest degree of accuracy or the least MSE value, 0.00025248.

RESEARCH METHODOLOGY

Literature of Research

A great deal of research has been conducted, some of which is obviously connected to the topic at hand. Before deciding to apply the model, the researcher decides to investigate some scholarly literature on forecasting stock prices using machine learning and deep learning. The researcher estimates the model to be utilized and the evaluation to be applied in this stage.

Data Preprocessing and Analysis

Data Collection

The material was collected from https://finance.yahoo.com (Yahoo Finance). The dataset utilized relies on three different companies:

1) BAC (Bank of America Corporation)
2) HDB (HDFC Bank Limited)
3) RY (Royal Bank of Canada)

Data Selection

Data selection is a phase that involves picking and removing usable and unnecessary data from the source. The data variable provided by the firm is as follows:

1) Date : The day on which trading on a newly issued stock begins.
2) Open : The first cost was paid on a day.
3) High : The biggest cost was paid on a day.
4) Low : The smallest cost was paid on a day.
5) Close : A closing cost was paid on a day.
6) Adj Close : A closing costs after dividends and stock splits have been deducted.
7) Volume : The total amount of deals that day.
8) HL_PCT : The percentages of the biggest price and smallest price for each day.
9) PCT_change : The percentages of first price and closing price for each day.
10) delta_open_close_day_before_% : The disparity between the closing price and the open next day's first price.
11) Open:30 days rolling : The mean of the preceding 30 days open price.
12) High:30 days rolling : A mean income of the preceding 30 days' high price.
13) Low:30 days rolling : A mean income of the preceding 30 days' low price.
14) Close:30 days rolling : A mean income of the preceding 30 days' close price.
15) Adj Close:30 days rolling : A mean income of the preceding 30 days' adj close price.

16) Volume:30 days rolling : A mean income of the preceding 30 days' volume price.
17) Label : Forecast out results

Only Adj Close, Volume, HL_PCT, and PCT_change were utilized from all variables in the study given.

Split Data

Before the data is divided into several training data and test data, the researcher determines how many adj close rows will be shifted to predict future stock prices, data that is not included in this number will be used as training data. After that, the researcher conducted 5 experiments. For the first experiment, researchers used 80% of the data that was not included in the shift rate. For the second experiment, researchers used 60% of the data that was not included in the shift rate. For the third experiment, researchers used 40% of the data that was not included in the shift rate. For the fourth experiment, researchers used 20% of the data that was not included in the shift rate. For the last experiment, researchers used 50% of the data that was not included in the shift rate.

Feature Scaling

Many machine learning or deep learning algorithms that measure convenience using Euclidean distance will fail to give satisfactory recognition for smaller features. Scaling crucial data in deep learning or machine learning models can be more effective. There are several ways for doing feature scaling, normalization, and standardization.

A) Standardization

Standardization ensures that all characteristics are planned in reference to an average value with a one-standard-deviation standard deviation. By decreasing the mean of each observation divided by the standard deviation it will reach standardization (1).

$$X_{new} = \frac{X - X_{mean}}{\sigma} \tag{1}$$

B) Normalization

Each value in a feature is reduced by the feature's minimum value, then divided by the range of values or the maximum value, and the decreased minimum value of the feature produces a new normalized value between 0 and 1 or -1 and 1 (2).

$$X_{new} = \frac{X_{old} - X_{min}}{X_{max} - X_{min}} \tag{2}$$

Models Evaluation

MAE, RMSE, and MAPE will be used to assess system performance. (1) MAE is defined as a mean absolute difference between measured and predicted values. (2) The RMSE is then computed by squaring the error (prediction) divided by the amount of data (= average), and finally rooted. (3) While MAPE is the absolute percentage of average error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \tag{2}$$

$$MAPE = \sum_{t=1}^{n}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| x\ 100\% \tag{3}$$

Literature of Research

The researcher employs a regression approach with five models such as LSTM, CNN, linear regression, lasso regression, and LSSVM. These model's performance will be assessed using numerous computations explained in the preceding section. Linear regression in quantitative research is to forecast the connection between variables X and Y. The reason the authors use linear regression is that despite its limitations, such as the fact that real data rarely shows a clear relationship between the dependent and independent variables, this linear regression can predict future values, making it useful for forecasting stock prices, sales, etc.

Lasso regression is an extension of linear regression in that the regularization parameters are multiplied by the sum of the absolute values of the weights and applied to the linear regression loss function (ordinary least squares). The gain using Lasso regression versus linear least squares regression is found in the bias-variance trade-off, which states that as alpha grows, the flexibility of the lasso regression's fit diminishes, resulting in a drop in variance but an increase in bias. The reasons why researchers utilize lasso regression to assist prevent overfitting since it has the capacity to set the coefficients for characteristics that are regarded unappealing to 0, hence lowering the model's complexity.

LSSVM is a variant of the standard SVM that employs equality constraints rather than inequality constraints and a squared loss function rather than the -insensitive loss function. LSTM is an RNN modification. The LSTM has three gates: A forget gate defines what information from the previous cell is to be disregarded, an input gate determines which data is aligned with the in-use cell, and an output gate determines which information should be transferred to the next hidden layer. The LSTM algorithm has the benefit of accepting input of varying durations. This capability is extremely handy for creating forecasting models with LSTM.
CNN is a subset of deep-neural-network. CNN is a technique for predicting future occurrences that uses a 1D array as input. CNN also has the benefit of being stronger at pattern recognition, more accurate at feature extraction, and faster at training. So that is the reason this research uses CNN

Design and Implementation

Design

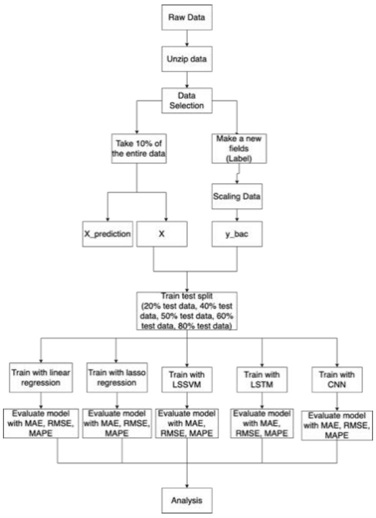
Raw Data
Unzip data
Data Selection
Take 10% of the entire data
Make a new fields (Label)
Scaling Data
X_prediction
X
y_bac
Train test split (20% test data, 40% test data, 50% test data, 60% test data, 80% test data)
Train with linear regression
Train with lasso regression
Train with LSSVM
Train with LSTM
Train with CNN
Evaluate model with MAE, RMSE, MAPE
Evaluate model with MAE, RMSE, MAPE
Evaluate model with MAE, RMSE, MAPE
Evaluate model with MAE, RMSE, MAPE
Evaluate model with MAE, RMSE, MAPE
Analysis

Figure 1. Methodology Flowchart

0 shows the flow of the proposed method applied in a single dataset. The flowchart is repeated three times with the same distribution of data for scenarios like those analyzed in this study, which contains three datasets. Following training and evaluation, comparison and analysis are performed. This stage informs the researcher on the most effective and correct model for this research.

Implementation

Preprocessing raw data is the first stage in constructing deep-learning and machine-learning approaches. The data preprocessing stage is when raw data is processed before being fed into machine-learning or deep-learning models.

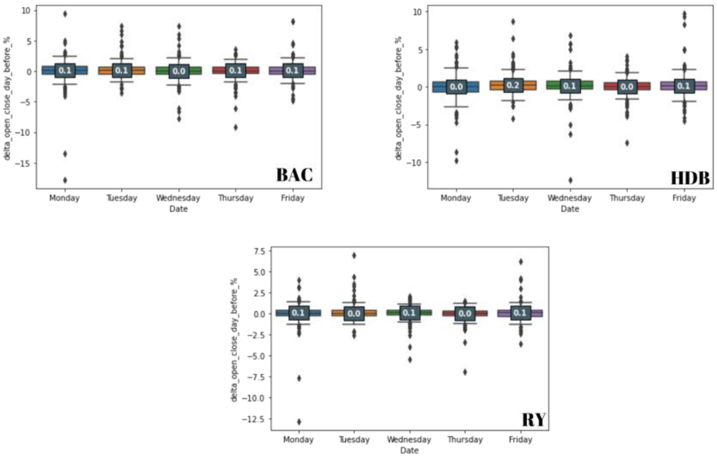

Figure 2. Gap between close price and open price

From 0 depicts the difference between the close price and the open price the following day, categorized by day. As seen in the graphic above, the BAC dataset has the greatest value, 0.1, on Monday, Tuesday, Thursday, and Friday, implying that selling stocks at the start of that day may be more beneficial. Whereas Tuesdays may be more profitable days to sell shares in the HDB dataset. Mondays, Wednesdays, and Fridays may be more profitable in the RY dataset.

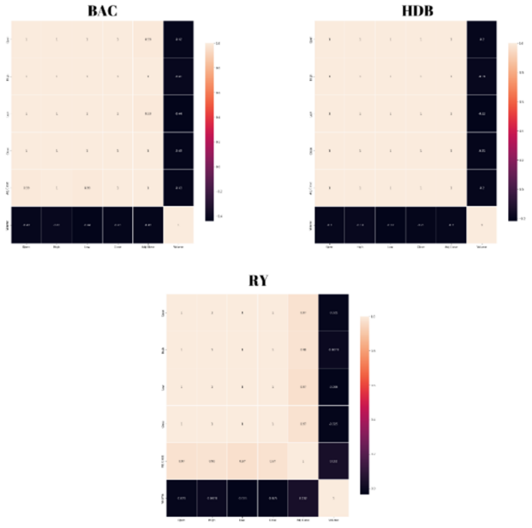

Figure 3. Heatmap of Dataset

0 shows how each variable is connected with one another; a score of 0 implies no correlation, whereas a score of 1 implies variable has a substantial positive link. This implies that when the value grows, so do the other values with other variables, however a negative value indicates that the value is negatively correlated with other values, which means that as the variable value increases, so do the other variable values since the correlation is negative.

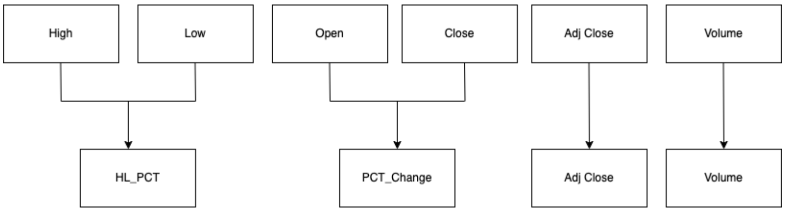

Figure 4. Data Selection

Looking at the previous section's heatmap data, we can observe that Open, High, Low, Close, and Adj Close have the strongest correlation. As a result, the researcher devised two new variables: HL PCT (the proportion of high and low prices per day) and PCT Change (the percentage of open and closed prices per day). As a result, the data will look like 0.

After that, we took 10% of the data in each dataset from the selected data then added a new variable that is utilized to move the adj closing price as much as the dataset is taken. Following that, we scaled the data that we previously created because the preceding data will be inserted into the dataset in the same range.

After successfully scaling the data, the data to be predicted is picked from 10% of the dataset, with the remainder being X data and the Y data containing array data from the adj closing price shift. The X and Y data will then be separated into five parts as explained in the previous sections and repeated it for each dataset.

Deep-Learning Models

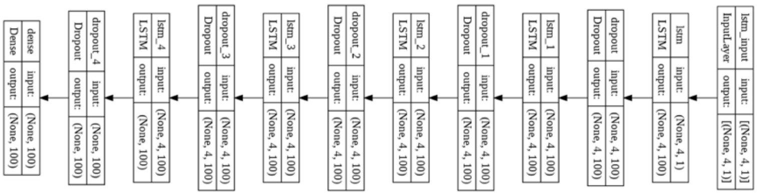

Figure 5. LSTM Model

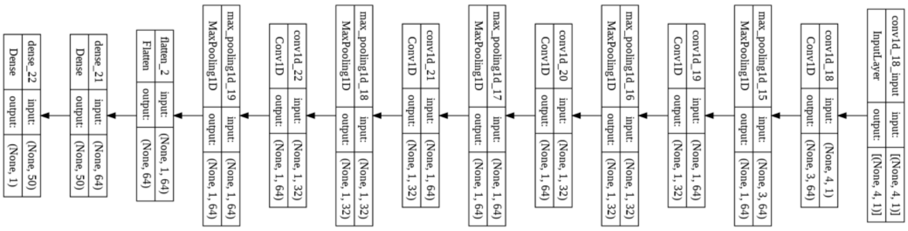

Figure 6. CNN Model

0 shows that there are 5 LSTM layers for one model. It begins with the value "[(None, 4, 1)]" as input. This number 4 is derived from the value of the initial X train shape array, and the number of variables is steadily raised until an output with 100 neurons. This is performed for the remaining two datasets by separating the training and test data portions as indicated in the previous section.

This 0 shows that the CNN model contains 5 layers as well. The input value is the same as the LSTM, "[(None,4,1)]," but the output value is different, namely if the CNN has 50 neurons.

Results and Analysis

Results

Following numerous training sessions with varying distributions of training data, the researcher acquired some assessment data, which will be discussed in the following step. Results of BAC can be seen on 0, results of HDB can be seen on 0, results of CNN can be seen on 0.

Table 1. Results of BAC

Testing		Models				
		Linear Regression	*Lasso Regression*	*LSSVM*	*LSTM*	*CNN*
2:8	***MAE***	4,424	4,519	3,738	4,901	3,744
	RMSE	5,540	5,765	5,840	6,377	4,866
	MAPE	15,234	15,450	13,341	16,511	13,146
4:6	***MAE***	4,243	4,452	3,390	5,638	3,378
	RMSE	5,323	5,612	4,935	6,531	4,549
	MAPE	15,255	15,889	12,279	20,461	11,954
5:5	***MAE***	4,165	4,369	3,417	4,483	3,617
	RMSE	5,290	5,558	4,900	6,128	4,864
	MAPE	14,942	15,556	12,351	14,997	12,578
6:4	***MAE***	4,155	4,344	3,405	3,399	3,390
	RMSE	5,252	5,470	4,927	4,765	4,623

Testing		Models				
		Linear Regression	*Lasso Regression*	*LSSVM*	*LSTM*	*CNN*
	MAPE	14,981	15,534	12,396	11,077	11,661
8:2	***MAE***	3,887	4,134	3,300	3,288	3,063
	RMSE	4,973	5,227	4,526	4,414	4,238
	MAPE	14,331	15,147	12,165	11,872	11,166

Table 2. Results of HDB

Testing		Models				
		Linear Regression	***Lasso Regression***	***LSSVM***	***LSTM***	***CNN***
2:8	***MAE***	7,488	8,372	7,662	7,499	7,203
	RMSE	10,084	10,594	10,691	9,950	9,838
	MAPE	13,395	14,894	13,662	13,182	12,713
4:6	***MAE***	7,532	8,115	7,608	7,535	7,732
	RMSE	9,954	10,335	10,080	9,842	9,836
	MAPE	13,955	15,034	14,016	13,838	14,594
5:5	***MAE***	7,472	8,080	7,585	7,924	7,095
	RMSE	9,970	10,400	10,280	9,968	9,666
	MAPE	13,766	14,849	13,967	15,229	12,480
6:4	***MAE***	7,513	8,083	7,478	7,389	7,537
	RMSE	10,039	10,408	10,000	9,720	9,622
	MAPE	13,805	14,750	13,731	13,454	14,113
8:2	***MAE***	7,180	7,838	7,252	6,595	6,751
	RMSE	9,700	10,137	9,498	8,865	9,002
	MAPE	13,195	14,450	13,407	11,810	12,271

Table 3. Results of RY

Testing		Models				
		Linear Regression	*Lasso Regression*	*LSSVM*	*LSTM*	*CNN*
2:8	***MAE***	7,269	8,517	5,332	8,737	5,131
	RMSE	9,541	10,946	7,435	12,230	7,139
	MAPE	9,950	11,507	7,488	11,316	7,209
4:6	***MAE***	6,960	8,120	5,310	5,474	6,097

Testing		Models				
		Linear Regression	*Lasso Regression*	*LSSVM*	*LSTM*	*CNN*
	RMSE	9,047	10,093	7,489	7,687	8,383
	MAPE	9,891	11,402	7,658	7,849	8,832
5:5	***MAE***	6,841	7,884	5,168	5,956	4,567
	RMSE	8,985	9,926	7,208	8,269	6,652
	MAPE	9,648	11,021	7,407	8,054	6,395
6:4	***MAE***	6,898	7,892	5,170	6,412	4,614
	RMSE	9,016	9,912	7,064	8,366	6,680
	MAPE	9,764	11,066	7,409	9,187	6,542
8:2	***MAE***	6,422	7,541	4,765	4,476	4,093
	RMSE	8,458	9,355	6,333	6,025	5,584
	MAPE	9,183	10,766	6,909	6,618	5,993

Analysis

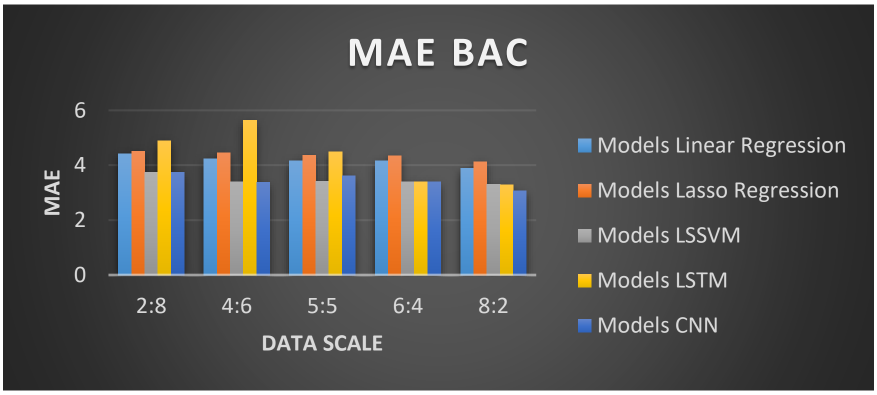

Figure 7. Mean Absolute Error of BAC

0 depicts the BAC dataset's Mean Absolute Error statistics. We can see from the graph above that the majority of CNN models have lower error levels. However, the error levels of the LSSVM and CNN models are essentially the same under some scenarios, such as the 2:8 scale. Overall, we may conclude that the CNN model is more successful than other models in this case.

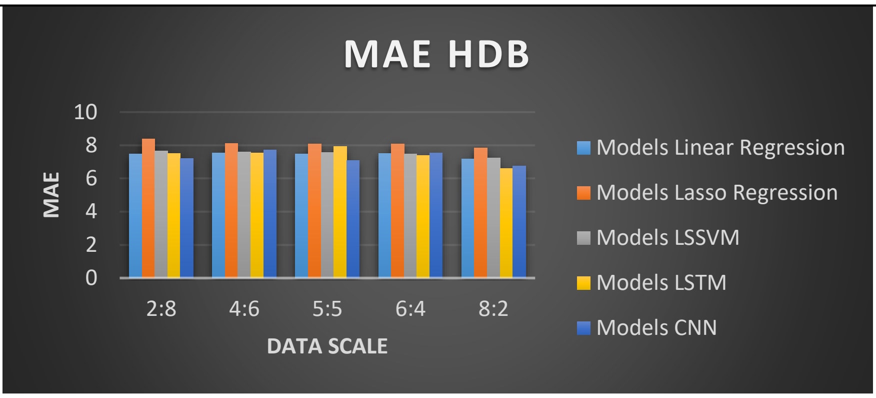

Figure 8. Mean Absolute Error of HDB

The Mean Absolute Error data from the HDB dataset is shown in 0. According to the graph above, the error levels of the Linear Regression and LSTM models are about the same under certain parameters, such as the 4:6 scale. Also, we can see from the graph that the LSTM and CNN models have about the same error numbers, but when we look at it again, the LSTM model is more effective than the CNN model.

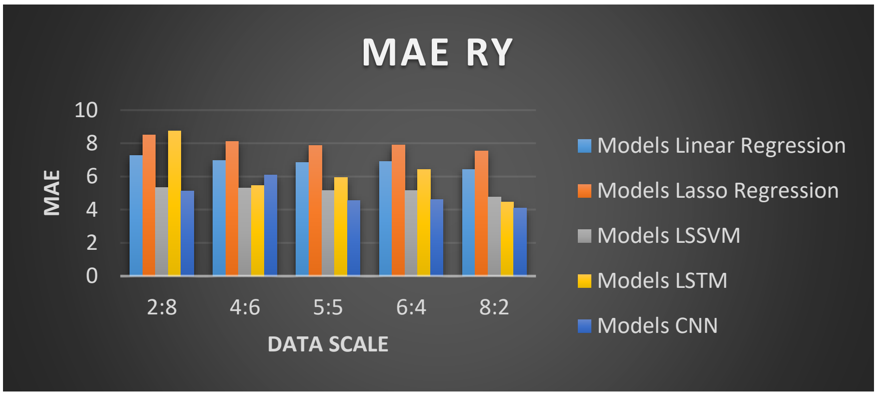

Figure 9. Mean Absolute Error of RY

The Mean Absolute Error data from the RY dataset is shown in 0. According to the graph above, the CNN model has a reduced error value for forecasting at all data sizes. Overall, we may conclude that the CNN model is more effective than the other models.

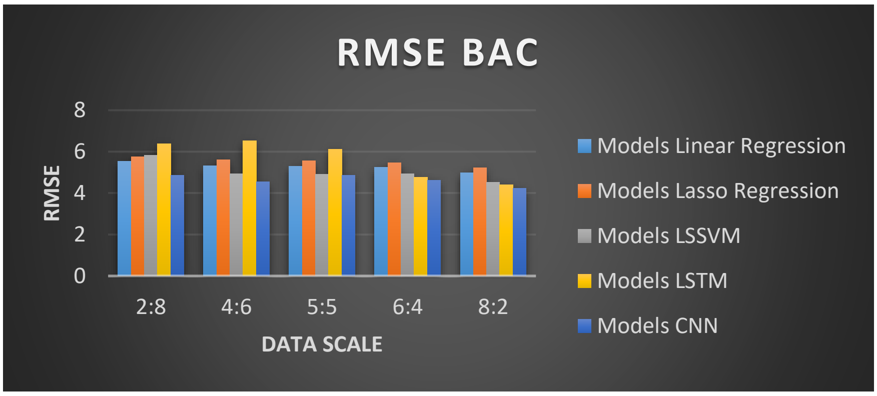

Figure 10. Root Mean Squared Error of BAC

The Root Mean Squared Error data from the BAC dataset is shown in 0. According to the graph above, the CNN model has a reduced error value for forecasting at all data sizes. Overall, we may conclude that the CNN model is more effective than the other models.

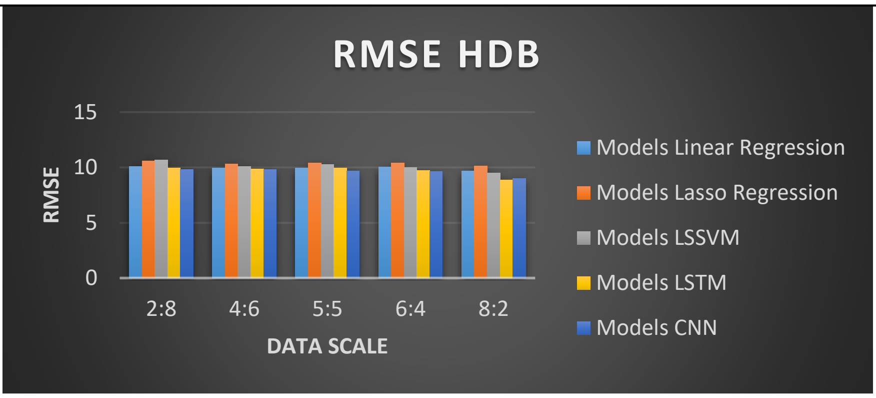

Figure 11. Root Mean Squared Error of HDB

0 depicts the HDB dataset's Root Mean Squared Error statistics. We can see from the graph above that the majority of CNN models have lower error levels. However, the error levels of the LSTM and CNN models are essentially the same under some scenarios, such as the 4:6 scale. Overall, we may conclude that the CNN model is more successful than other models in this case.

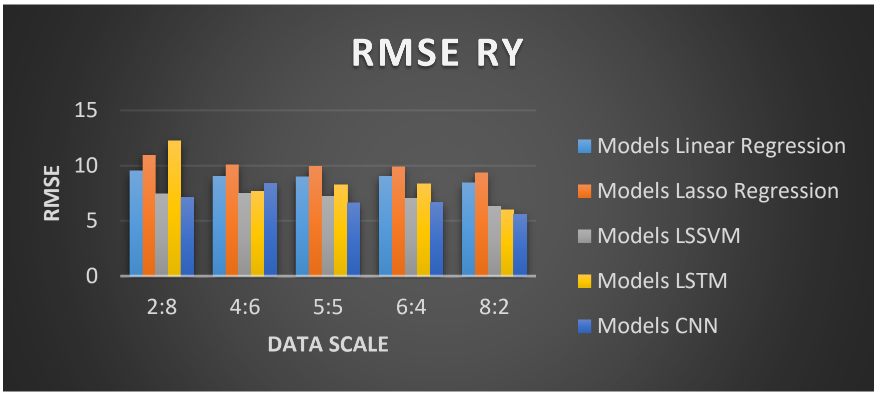

Figure 12. Root Mean Squared Error of RY

The Root Mean Squared Error data from the RY dataset is shown in 0. According to the graph above, the CNN model has a reduced error value for forecasting at all data sizes. Overall, we may conclude that the CNN model is more effective than the other models.

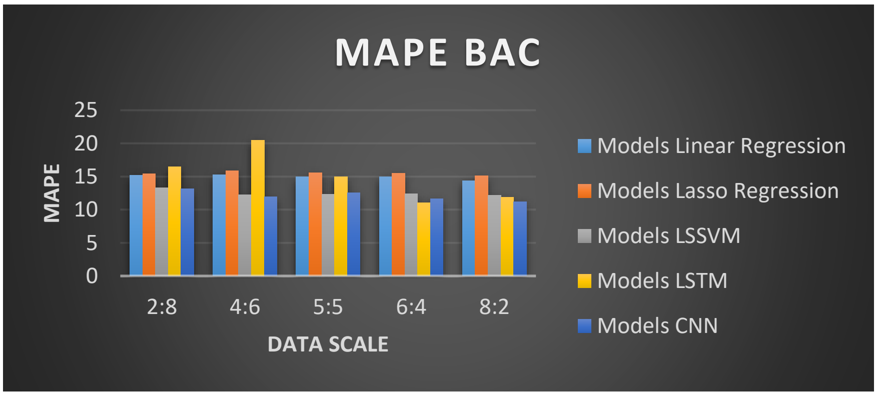

Figure 13. Mean Absolute Percentage Error of BAC

The Mean Absolute Percentage Error data from the BAC dataset is shown in 0. Based on the graph above, at a 5:5 data scale, the LSSVM and CNN models have almost the same error values. However, the majority CNN model has a lower error value so that the CNN model is more effective in forecasting.

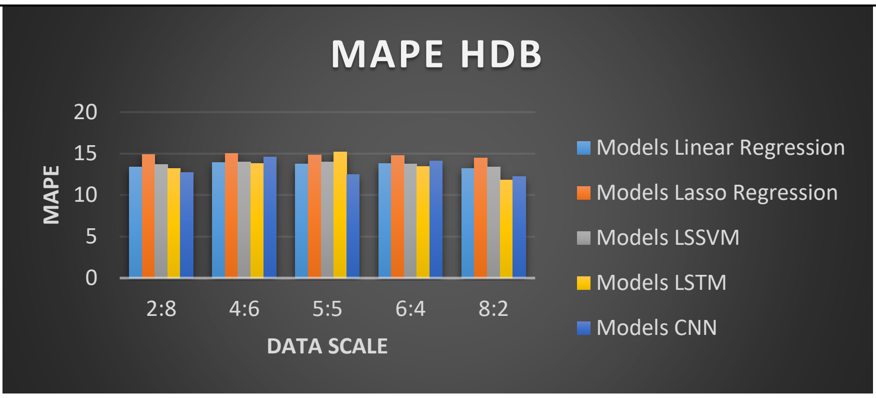

Figure 14. Mean Absolute Percentage of HDB

The Mean Absolute Percentage Error data from the HDB dataset is shown in 0. Based on the graph above, at a 4:6 data scale, the Linear Regression and LSTM models have almost the same error values. However, the majority LSTM model has a lower error value so that the LSTM model is more effective in forecasting.

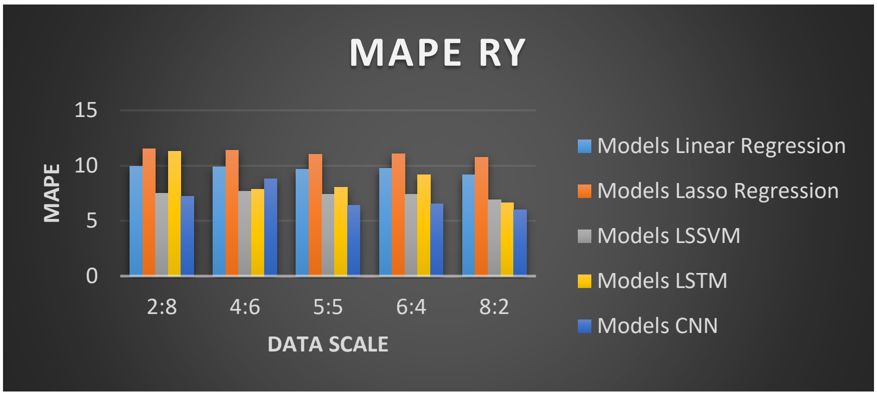

Figure 15. Mean Absolute Percentage of RY

The Mean Absolute Percentage Error data from the RY dataset is shown in 0. Based on the graph above, at a 4:6 data scale, the LSSVM and LSTM models have almost the same error values. However, the majority CNN model has a lower error value so that the CNN model is more effective in forecasting.

Conclusion

There are several methods and algorithms for forecasting stock prices based on the outcomes of the tests that have been conducted. An algorithm will produce predictable patterns in data, which will lead to improved tactics and more accurate forecasts. Predictive data must incorporate behavioral trends, previous transactions, and demographic information. Machine learning can forecast what proportion of transactions will occur in the future based on this data. This is accomplished by gathering historical data, training the model, and then adjusting the parameters to evaluate/apply the forecast model.

Based on several models that were trained and evaluated, the results obtained are as described in the previous section, seen in the BAC dataset, the CNN algorithm as a whole gets a lower error value than other algorithms, indicating that the CNN algorithm is more effective and accurate in this dataset with a value of The MAPE with the lowest value is 11,166. In the HDB dataset, the LSTM technique outperforms other models with the lowest MAPE value of 11,81. Finally, with the lowest MAPE value of 5,993 for the RY dataset, which is the same as the BAC dataset, the CNN method is more effective and accurate than the other algorithms. Researchers advise expanding the dataset and taking more exact measurements. It would be even better if the researcher used various methods to obtain lower error levels and improved accuracy in future study.

REFERENCES

[1] S. Mehtab and J. Sen, "A Time Series Analysis-Based Stock Price Prediction Using Machine Learning and Deep Learning Models," *IJBFMI*, vol. 6, no. 4, p. 272, 2020, doi: 10.1504/IJBFMI.2020.115691.

[2] B. Pavlyshenko, "Machine-Learning Models for Sales Time Series Forecasting," *Data*, vol. 4, no. 1, p. 15, Jan. 2019, doi: 10.3390/data4010015.

[3] S. Siami-Namini, N. Tavakoli, and A. Siami Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, Dec. 2018, pp. 1394–1401. doi: 10.1109/ICMLA.2018.00227.

[4] N. Milosevic, “Equity forecast: Predicting long term stock price movement using machine learning.” arXiv, Nov. 22, 2018. doi: 10.48550/arXiv.1603.00751.

[5] J. Sen, “Stock Price Prediction Using Machine Learning and Deep Learning Frameworks”, Accessed: Dec. 14, 2022. [nline]. Available: https://www.academia.edu/38029599/Stock_Price_Prediction_Using_Machine_Learning_and_Deep_Learning_Frameworks .

[6] E. Mussumeci and F. C. Coelho, “Machine-learning forecasting for Dengue epidemics - Comparing LSTM, Random Forest and Lasso regression,” Public and Global Health, preprint, Jan. 2020. doi: 10.1101/2020.01.23.20018556.

[7] B. Abdualgalil, S. Abraham, W. M. Ismael, and D. George, “Modeling and Forecasting Tuberculosis Cases Using Machine Learning and Deep Learning Approaches: A Comparative Study,” in Data Management, Analytics and Innovation, vol. 137, S. Goswami, I. S. Barara, A. Goje, C. Mohan, and A. M. Bruckstein, Eds. Singapore: Springer Nature Singapore, 2023, pp. 157–171. doi: 10.1007/978-981-19-2600-6_11.

[8] S. Mehtab, J. Sen, and A. Dutta, “Stock Price Prediction Using Machine Learning and LSTM-Based Deep Learning Models,” in Machine Learning and Metaheuristics Algorithms, and Applications, vol. 1366, S. M. Thampi, S. Piramuthu, K.-C. Li, S. Berretti, M. Wozniak, and D. Singh, Eds. Singapore: Springer Singapore, 2021, pp. 88–106. doi: 10.1007/978-981-16-0419-5_8.

[9] S. Ismail and A. Shabri, “Time Series Forecasting using Least Square Support Vector Machine for Canadian Lynx Data,” Jurnal Teknologi, vol. 70, no. 5, Sep. 2014, doi: 10.11113/jt.v70.3510.

[10] P. J. Bastos, “Thesis Title: Bear Market Prediction Using Logistic Regression, Random Forest, and XGBoost”, November. 2019, Available: https://fenix.tecnico.ulisboa.pt/downloadFile/1689244997259684/thesis80959pb.pdf.

[11] A. Triyono, R. B. Trianto, and D. M. P. Arum, “Penerapan Least Squares Support Vector Machines (LSSVM) dalam Peramalan Indonesia Composite Index,” *JIUP*, vol. 6, no. 1, p. 210, Mar. 2021, doi: 10.32493/informatika.v6i1.10237.