

Forecasting Next-Day TLKM Stock Prices Using Linear and Nonlinear Machine Learning Models: A Case Study of PT Telekomunikasi Indonesia Tbk

Edwin Setiawan Nugraha^{1*}, Azizah Dwi Aprilia², Samsul Arifin³ and Norani Saro⁴

^{1,2} Department of Actuarial Science, President University, Indonesia

³ Department of Data Science, Institut Teknologi Sains Bandung, Indonesia

⁴ Faculty of Humanities and Education, UNITAR International University, Malaysia

ABSTRACT

Open Access

Type: Research Paper

Article:

Received: January 25, 2026

Revised: March 16, 2026

Accepted: July 13, 2026

Issued: October 30, 2026

© Nugraha et al. (2026).



This is an open-access article
under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license

*Correspondence Address:

Edwin.nugraha@president.ac.id

Stock price forecasting provides useful information for investors in making decisions in fluctuating capital markets. The objective of this work is to develop and compare models to forecast the closing price for the next trading day. Three models were employed: linear regression, an optimized random forest regression, and an optimized backpropagation neural network. The dataset used consists of daily market data for TLKM and IHSG covering the period from March 22, 2018, to February 27, 2023. The seven predictors used are open price, high price, low price, trading volume, SMA5, RSI14, and IHSG return, while the target variable is the next day's TLKM closing price. The dataset was divided chronologically into 80% training and 20% testing sets. Hyperparameter optimization was performed with 5-fold time-series cross-validation, using RandomizedSearchCV with 50 configurations for the random forest and a random search with 10 configurations for the backpropagation neural network. The results show that linear regression outperforms the other models based on the overall evaluation metrics, with an MAE of 57.67, RMSE of 81.38, MAPE of 1.37%, and R^2 of 0.928. The MAPE value indicates that the linear regression model has a relatively low forecasting error. The linear regression model can also explain approximately 92.8% of the variation in the next day's TLKM closing price through the predictors included in the model. The implication of this work is that a linear regression model can provide an additional quantitative reference for investors in evaluating the possibility of TLKM closing price movements on the next trading day.

Keywords: closing price, linear regression, random forest, backpropagation, tlkm

1. INTRODUCTION

Investing in stocks offers the potential for higher returns compared with other financial assets, such as savings, time deposits, and bonds. However, investors also need to be aware of the risk of losses because stock prices fluctuate over time. The risk–return trade-off is a fundamental concept in investment, where higher expected returns are generally associated with greater investment risk [1]. When stock prices increase, investors may earn profits, whereas a decline in stock prices may result in losses.

Future stock price movements cannot be predicted with certainty because they are influenced by various factors, including macroeconomic conditions, company performance, market conditions, and investor behaviour. The ability to forecast future stock prices can provide useful information for investors and serve as an additional consideration in making investment decisions [2].

A review article identified several predictors in market forecasting, including historical prices, trading variables, technical indicators, and market-related variables [2]. Other studies [3], [4] utilized historical prices and trading variables, specifically Open, High, Low, and Volume, as predictors to model stock prices using multiple linear regression. Additionally, technical factors such as Simple Moving Average (SMA) and Relative Strength Index (RSI) have been employed to model price trends and momentum [2], [5]. Market condition factors, such as the Jakarta Composite Index (IHSG), have also been used as variables in stock price modeling [6]. Based on the variables utilized in these previous studies, this work considers using seven predictors, including Open, High, Low, Volume, SMA₅, RSI₁₄, and IHSG return.

In building a forecasting model, one important consideration is the relationship between predictors and stock prices. Due to market complexity and dynamics, this relationship is often intricate. Evaluating models with different approaches, such as linear and non-linear models, is crucial. Studies in [7], [8] compared Linear Regression (LR), Random Forest (RF), and neural network models for stock price prediction. In the Indonesian market, LR, RF, and neural network models

were compared for predicting the TLKM closing price [9]. Another study presented a comparison among LR, RF, and XGBoost for TLKM stock price prediction [10]. Recent research has also applied CNN, LSTM, BiLSTM, and hybrid CNN–BiLSTM models to short-term stock price forecasting using historical market data and financial news sentiment [11]. Other recent studies have investigated optimized LSTM models for stock price forecasting [12] and hybrid GRU–Transformer architectures for capturing temporal patterns in stock price data [13]. More recently, Transformer-based architectures have also been developed to model complex temporal and cross-stock relationships in stock price forecasting [14]. Systematic and bibliometric reviews further indicate the growing application of machine learning, ensemble methods, and artificial intelligence techniques in financial market prediction [15], [16].

These studies demonstrate the applicability of different levels of model complexity to stock price forecasting, while also indicating that model performance may depend on the data, predictors, forecasting task, and model configuration.

In addition to the relationship between predictors and stock prices, preserving temporal information during model development and evaluation must also be considered. Financial observations are chronologically ordered, and a forecasting model should only use information available at the prediction time when estimating future observations. A study in [17] presents a discussion on the use of cross-validation for time-series forecasting and emphasizes that the validity of the validation procedure depends on the characteristics of the forecasting problem. This study preserves the chronological order of observations when dividing the data into training and test sets and employs time-series cross-validation for model development and hyperparameter selection. Although previous studies have investigated several machine learning models for TLKM [9], [10], a controlled comparison of linear and nonlinear models using the same predictor set, a clearly defined next-day forecasting target, chronological data partitioning, and time-series-based hyperparameter optimization remains highly relevant.

In this study, we compare the

forecasting performance of Linear Regression (LR), Random Forest (RF), and Backpropagation Neural Network (BPNN) models for predicting the next-day closing price of TLKM. Based on the predictors identified in [2]–[6], we selected seven predictors: Open, High, Low, Volume, SMA₅, RSI₁₄, and IHSG Return. The data were chronologically partitioned into training and test sets. For RF and BPNN, hyperparameter optimization was conducted using five-fold Time-Series Cross-Validation. Specifically, RandomizedSearchCV with 50 randomly selected configurations was applied to RF, whereas a Random Search with 10 configurations was applied to BPNN. The final

models were evaluated on an unseen test set using MAE, MSE, RMSE, MAPE, and R². Through this framework, this study provides a controlled empirical comparison between linear and nonlinear approaches for next-day TLKM closing-price forecasting.

2. METHODS

2.1 Research Framework

This research workflow is presented in Figure 1. The first step is data preparation, which includes the collection and preprocessing of daily TLKM and IHSG data before the models are implemented.

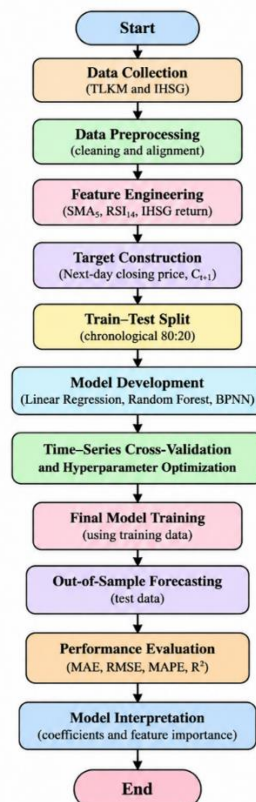


Figure 1. Research Framework

The next step is feature engineering, which constructs new variables, namely SMA₅, RSI₁₄, and IHSG Return, as well as the target variable, the next-day closing price. The data are then split chronologically into training and testing sets. Next, three models are implemented: Linear Regression (LR), Random Forest Regression (RF), and Backpropagation Neural Network (BPNN). During model development, five-fold time-series cross-validation and hyperparameter optimization are

performed for RF and BPNN.

2.2 Data Source

This study uses daily market data for TLKM obtained from Yahoo Finance through web scraping for the period from March 22, 2018, to February 27, 2023. The dataset contains the open, high, low, and close prices, the trading volume, and the IHSG index, with 1,203 observations. Each row represents one trading day of TLKM. To include technical

information, feature engineering was performed to create the SMA5 and RSI14 predictors. In addition, the IHSG return predictor was created to capture the effect of market movements. The target variable is defined as the next-day closing price.

2.3 Feature Engineering and Target Definition

We employ feature engineering to construct the SMA5, RSI14, and IHSG return predictors, which are briefly explained below. SMA₅ refers to the Simple Moving Average of closing price over the last five trading days, serving to smooth short-term price fluctuations and identify trend direction. Using a five-day trading window allows SMA₅ to represent short-term price tendencies, where the price position relative to the SMA₅ indicates short-term trend direction and momentum. The formula is expressed as follows:

$$SMA_{5,t} = \frac{1}{5} \sum_{i=1}^5 Close_{t-i} \quad (1)$$

Next, RSI14 refers to the 14-day Relative Strength Index, a momentum indicator measuring the speed and magnitude of price changes over the last 14 trading periods. Ranging from 0 to 100, values above 70 indicate overbought conditions, whereas values below 30 signal oversold conditions. These levels indicate potential price exhaustion or reversal, though such signals may not guarantee a trend reversal during strong trends. The formula of RSI₁₄ is given by

$$RSI_{14,t} = 100 - \frac{100}{5 + RS_t} \quad (2)$$

Where

$$RS_t = \frac{\text{average gain}_{14}}{\text{average loss}_{14}} \quad (3)$$

The last one, IHSG Return is used as an indicator of aggregate stock market movements and also represents general market conditions. Movements in the IHSG can provide information regarding broader market directions that are potentially related to individual stock movements. By incorporating IHSG Return as a predictor, the model obtains additional information regarding overall market conditions beyond the price characteristics and technical indicators of TLKM. IHSG daily

return is formulated by

$$R_{IHSG} = \frac{IHSG_t}{IHSG_{t-1}} - 1 \quad (4)$$

2.4 Data Processing

In this step, the data are preprocessed before model development. This step is important to ensure proper temporal alignment and consistent data splitting and to avoid information leakage. The process consists of three stages: 1) missing-value checking and data alignment, 2) chronological train-test split, and 3) feature scaling. In the first stage, the original dataset was checked for missing values, and none were identified. The target variable was constructed by shifting the TLKM closing price one trading day forward. If $X_t = \{X_{1,t}, X_{2,t}, X_{3,t}, X_{4,t}, X_{5,t}, X_{6,t}, X_{7,t}\}$, we can define that Y_{t+1} is target variable as a function of predictor X_t

$$Y_{t+1} = f(X_t) \quad (5)$$

where $X_{1,t}$ represent the open price at time t , $X_{2,t}$ represent the high price at time t , $X_{3,t}$ represent the low price at time t , $X_{4,t}$ represent the volume at time t , $X_{5,t}$ represent the SMA₅ at time t , $X_{6,t}$ represent the RSI14 at time t , $X_{7,t}$ represent the IHSG return at time t . These notations are used throughout the remainder of this work. Note that the last observation is removed because it has no next-day closing price. In the second stage, a chronological train-test split is applied because the dataset is a time series and therefore cannot be split randomly. This means that the training data chronologically precede the testing data. The 1,202 observations are divided into 80% training data and 20% testing data. The training data are used for model development, cross-validation, and hyperparameter optimization, while the testing data are used to evaluate model performance. In the third stage, to support efficient data processing for BPNN, the original data are normalized using the formula

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (6)$$

where x is the variable's value, x_{min} and x_{max} are the minimum and maximum values, respectively, obtained from the training data. Data leakage is prevented by not using the testing data for model training, cross-validation, or hyperparameter optimization. In each cross-

validation fold, the scaler is fitted on the training portion and then applied to normalize the validation portion of that fold.

2.5 Forecasting Models

We implement three models to forecast the next-day closing price of TLKM: Linear Regression (LR), Random Forest (RF), and Backpropagation Neural Network (BPNN). All models use the same predictor variables and forecasting target as given in equation (5).

2.5.1 Linear Regression

This model assumes that the predictor variables at time t have a linear relationship with the target variable at time $t+1$. The model can be expressed as

$$Y_{t+1} = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \dots + \beta_7 X_{7,t} \quad (7)$$

Here, β_0 is intercept, β_i are the regression coefficients associated with the predictor variables. These coefficients are estimated using ordinary least squares on the training data [18], [19]. This model was selected as the benchmark because of its simple and interpretable structure.

2.5.2 Random Forest

The second model is Random Forest, a nonlinear ensemble learning model consisting of multiple decision trees. Each tree is trained on a bootstrap sample of the training data using a randomly selected subset of features [18], [19]. The final prediction is obtained by averaging the outputs of all trees, as formulated below:

$$\hat{y}_{t+1} = \frac{1}{N} \sum_{i=1}^N T_i(X_t) \quad (8)$$

where $T_i(X_t)$ represents the output of the i -th tree and X_t represents the set of predictor variables. In addition, this model provides feature importance, which measures the contribution of each predictor variable to the prediction of the target variable [18].

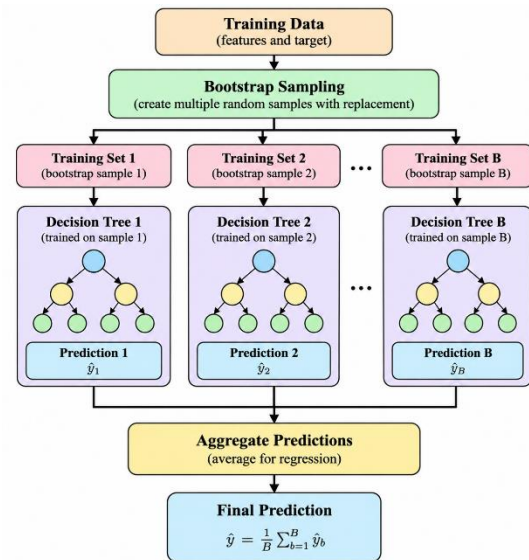


Figure 2. Simplified Illustration of Random Forest Regression Process

This feature importance is given by

$$I(X_i, t) = \frac{1}{N} \sum_{i=1}^N \sum_{m \in S_b(k)} \Delta mse(m) \quad (9)$$

where $\Delta mse(m)$ is the reduction in error at node m . The higher the value of $I(X_i, t)$, the greater the contribution of the predictor variable. The overall process of forecasting the next-day TLKM closing price using the random forest model is summarized in Figure 2.

2.5.3 Backpropagation Neural Network

The other nonlinear model employed in this work is the Backpropagation Neural Network (BPNN). Its architecture consists of three layers: an input layer, a hidden layer, and an output layer. Data are transformed from the input layer to the hidden layer through an activation function. Common activation functions include the linear, sigmoid, hyperbolic tangent, and ReLU functions [18], [19]. Let f be the activation function; the hidden-layer output can be written as follows:

$$h_{ij} = f(\sum_i^N w_{ij} x_i + b_j) \quad (10)$$

where w_{ij} represents the weight connecting the i -th input to the j -th neuron in the hidden layer, and b_j is the corresponding bias. The hidden layer uses the hyperbolic tangent ($\tanh$) activation function to introduce nonlinearity, while a linear activation function is used in the output layer because the target variable is a continuous stock closing price.

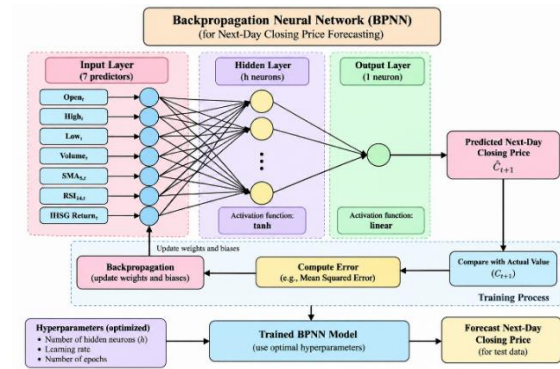


Figure 3. BPNN architecture for next day TLKM price forecasting

The number of input neurons is set to seven, corresponding to the seven predictor variables. Model training consists of three steps. The first step is the forward propagation process in which the input is propagated in the network to produce the output. The second step is backward propagation, which starts with error calculation and applies gradient descent to determine the direction and magnitude of weight and bias updates. In the last step, the weights and biases are updated using the formula

$$w^{new} = w^{old} - \eta \frac{\partial L}{\partial w} \quad (11)$$

here, η is the learning rate, and $\frac{\partial L}{\partial w}$ is the gradient of the loss function with respect to the weight. These steps are repeated iteratively until the number of epochs determined by hyperparameter optimization is reached. The BPNN architecture used to forecast the next-day TLKM closing price, with seven input variables, one hidden layer, and a single output, is illustrated in Figure 3.

2.6 Implementation and Reproducibility

To facilitate reproducibility, we present the software environment and computational settings used in this study, as shown in Table 1. These details include the programming environment, software libraries, random seed, and key BPNN implementation settings.

Table 1. Implementation and Reproducibility Detail

Component	Specification
Programming language	Python 3.11.0
NumPy	2.4.6
pandas	3.0.3
scikit-learn	1.9.0

Component	Specification
Matplotlib	3.11.0
Random seed	42
Data normalization	Min-Max Scaling
BPNN hidden-layer activation	Hyperbolic tangent (tanh)
BPNN output-layer activation	Linear
BPNN weight initialization	Xavier-style initialization
BPNN bias initialization	Zero initialization
BPNN optimization	Manual gradient descent
BPNN hyperparameter optimization	Random Search (10 trials)

2.7 Time Series Cross Validation

To ensure that model development and hyperparameter optimization follow the chronological order of stock prices, we employed time-series cross-validation. Figure 4 shows the five-fold time-series cross-validation used in this study. In each fold, the model is trained on data from past periods and tested on the subsequent period.

As validation moves forward in time, the training set gradually expands with newer data. This ensures that the model never uses future information to predict past observations.

2.8 Hyperparameter Optimization

This section describes hyperparameter optimization for the random forest and backpropagation neural network. For the random forest, we use RandomizedSearchCV with 50 randomly selected hyperparameter combinations.

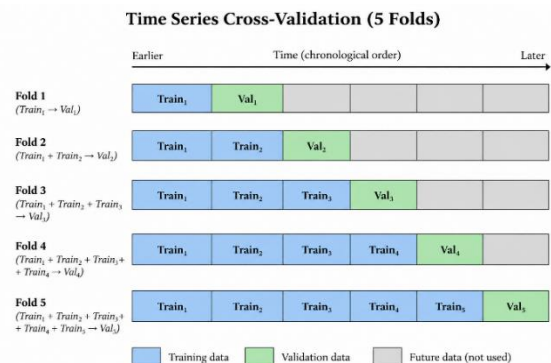


Figure 4. Five-Fold time series cross validation for model development and hyperparameter optimization

This search is implemented using five-fold time-series cross-validation, with RMSE as the evaluation metric. The hyperparameters considered are the number of trees ($n_estimators$), maximum tree depth

(max_depth), minimum number of samples required to split a node (min_samples_split), minimum number of samples at a leaf node (min_samples_leaf), and the number of features considered at each split (max_features). The optimal configuration is selected based on the lowest mean cross-validation RMSE. For the backpropagation neural network, we apply a random search with 10 randomly selected configurations, each evaluated using five-fold time-series cross-validation. As with the random forest, the configuration with the lowest mean cross-validation RMSE is selected. The hyperparameter search space for both models is presented in Table 2.

Table 2. Hyperparameter Search Space

Model	Variable	Search Space
RF	n_estimators	100, 200, 300, 500, 700, 1000
	max_depth	None, 5, 10, 15, 20, 25, 30
	min_samples_split	2, 5, 10, 15, 20
	min_samples_leaf	1, 2, 4, 8, 10
	max_features	1.0, sqrt, log2, 0.7, 0.8
BPNN	hidden neurons	1–10
	learning rate	10^{-5} , 5×10^{-5} , 10^{-4} , 2×10^{-4} , 5×10^{-4}
	epochs	10,000–50,000

2.9 Performance Evaluation

This section describes the performance evaluation of the three forecasting models. The evaluation aims to measure model's predictive errors and model's ability to explain the variation of the next day TLKM closing price. We employ five evaluation metrics: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and the coefficient of determination (R^2).

Let n be the number of observations, y_i represents the actual value and $\hat{y}_i$ represents the predicted closing price. The evaluation metrics used in this study are presented below.

MAE measures the average absolute difference between the actual and predicted values. A lower MAE indicates a smaller prediction error. The formula of MAE is given by

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}_i| \quad (12)$$

MSE measures the average squared difference between the actual and predicted values. This measure is sensitive to large errors because of the squaring effect; a single large error can dominate and penalize the model's performance score far more than the accumulation of many small errors. The formula of MSE is expressed by

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (13)$$

RMSE indicates the average magnitude of prediction errors and has the same unit as the target variable. It is obtained by taking the square root of MSE as follows:

$$RSME = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (14)$$

Mathematically, MAPE is defined as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (15)$$

This measure represents the prediction error as a percentage of the actual value. A lower MAPE indicates more accurate predictions.

According to [20], MAPE values are classified into four criteria indicating the quality of prediction results, as shown in Table 3.

Table 3. Predicting criteria

MAPE	Predicting Capability
< 10%	Highly accurate predicting
10% - 20%	Good predicting
20% - 50%	Reasonable predicting
> 50%	Weak and inaccurate predicting

The last evaluation metric is the coefficient of determination, denoted by R^2 . It is calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (16)$$

This value represents the proportion of variation in the target variable explained by the model. A higher value indicates better explanatory performance.

3. RESULT AND DISCUSSION

This section discusses the results of the forecasting models for next-day TLKM closing prices, focusing on descriptive statistics, correlation and multicollinearity, hyperparameter optimization, out-of-sample forecasting, and model interpretation.

3.1 Descriptive Statistics

Table 4 presents the descriptive statistics of all variables in this study, both predictors and target, based on the dataset of 1,202 observations. The mean TLKM opening, high and low prices were 3,742.27, 3,786.78, and 3,697.18, respectively. The mean of next day close price was 3,742.08 with standard deviation of 471.42.

Table 4. Descriptive statistics of variables

	Mean	Std. Dev	Min	Median	Max
X1	3,742.27	469.45	2,550	3,780	4,850
X2	3,786.78	467.65	2,590	3,820	4,850
X3	3,697.18	469.52	2,450	3,740	4,720
X4	1.0962	0.6716	0	0.929	6.241
X5	3,741.27	467.65	2,598	3,780	4,696
X6	49.76	13.69	11.54	50.00	86.49
X7	0.000116	0.0108	-0.0658	0.0005	0.1019
Y	3,742.08	471.42	2,560	3,780	4,770

Note: The unit of X₄ is 10⁸ shares.

The momentum variable RSI14 has a mean of 49.76, with values ranging from 11.54 to 86.49, indicating substantial variation in momentum during the study period. The mean IHSG return is 0.000116, with a standard deviation of 0.0108. The mean daily IHSG return is close to zero, while its relatively large standard deviation indicates considerable day-to-day variability during the observation period. The standard deviation is approximately 93 times larger than the mean, showing that daily fluctuations in IHSG returns dominate the long-term growth trend. IHSG returns exhibited both positive and negative movements, ranging from a minimum of -0.0658 to a maximum of 0.1019. Trading volume exhibited substantial dispersion, with a mean of approximately 109.62 million shares and a maximum of 624.14 million shares.

3.2 Correlation and Multicollinearity

Correlation and multicollinearity analyses were employed to investigate the relationships among the predictors, namely TLKM prices, technical indicators, trading volume, and IHSG return. This analysis helps identify potential redundancy among the predictors.

Figure 5 shows the Pearson correlation matrix of the variables. According to this result, Open, High, Low, SMA5, and Close have very high positive correlations, whereas Volume, RSI14, and IHSG Return have relatively weak correlations with the price-related predictors.

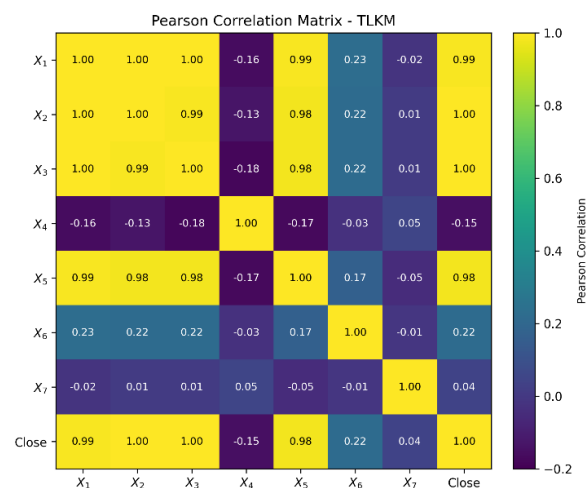


Figure 5. Pearson correlation matrix of the variables used in the forecasting model

The second analysis is the Variance Inflation Factor (VIF), whose results are shown in Table 5. The VIF values for Open, High, Low, and SMA5 are far greater than 10, indicating multicollinearity among these predictors.

Table 5. VIF of the Predictor Variables

Feature	VIF
X ₁	211.7815
X ₂	172.0872
X ₃	198.0321
X ₄	1.377094
X ₅	46.90003
X ₆	1.231911
X ₇	1.338172

The other predictors, namely Volume, RSI14, and IHSG Return, have VIF values below 5, indicating no multicollinearity. These results confirm that the price-related predictors overlap heavily. This makes it difficult to interpret individual variables in the linear regression model, whereas the random forest

and backpropagation neural network are less sensitive to this issue.

3.3 Cross Validation and Hyperparameter Optimization

This section presents the results of cross-validation and hyperparameter optimization for the RF and BPNN models. Using RandomizedSearchCV to optimize the RF hyperparameters across 50 randomly selected combinations via five-fold Time-Series Cross-Validation, we found the best configuration consisted of 1,000 trees, a maximum depth of 20, a minimum of 2 samples for splitting, a minimum of 1 sample per leaf, and all features considered at each split. This configuration achieved a mean cross-validation RMSE of 128.61. For the BPNN, we applied a random search across 10 randomly selected configurations using the same five-fold Time-Series Cross-Validation procedure. The best configuration used 4 hidden neurons, a learning rate of 0.0005, and 40,000 training epochs, achieving a mean cross-validation RMSE of 109.91.

3.4 Forecasting Performance

This section presents the test-set performance of LR, optimized RF, and optimized BPNN. The results show that LR achieves the lowest MAE, MSE, RMSE, and MAPE among the models. Additionally, LR obtains the highest R^2 value, as shown in Table 6. These results demonstrate that the nonlinear models, RF and BPNN, did not provide an improvement over LR for the given predictor variables and forecasting period. Thus, LR is the best-performing model for forecasting the next-day TLKM closing price on the test data.

Table 6. Forecasting performance on the test set

Model	MAE	MSE	RMSE	MAPE	R^2
LR	57.67	6,623.44	81.38	1.37	0.928
RF	91.12	15,190.14	123.25	2.10	0.835
BPNN	112.22	18,648.02	136.56	2.61	0.797

Next, we examine model generalization and stability by comparing the training and test performance of the three forecasting models, as shown in Table 7.

Table 7. Comparison of training and test performance

Model	Train RMSE	Test RMSE	Train R^2	Test R^2
LR	82.75	81.38	0.963	0.928

Model	Train RMSE	Test RMSE	Train R^2	Test R^2
RF	29.93	123.25	0.995	0.835
BPNN	130.75	136.56	0.908	0.797

This comparison provides additional insight into how consistently each model performs on unseen data. As shown in Table 7, LR shows relatively consistent performance, with its RMSE changing slightly from 82.75 on the training data to 81.38 on the testing data and R^2 shifting from 0.963 to 0.928. RF shows significant performance degradation in both RMSE and R^2 , whereas BPNN remains relatively stable but less accurate than LR and RF. This result highlights differences in generalization and stability across the models, with LR showing the most consistent performance.

3.5 Model Interpretation

This section identifies the predictors that contribute to the forecasting results and evaluates their contributions based on LR coefficients and RF feature importance. Table 8 shows that Open, High, Low, and SMA5 have positive LR coefficients of similar magnitude. This indicates that these price-related predictors are positively associated with the next-day TLKM closing price. RS114 also has a positive coefficient, but its value is very small.

Table 8. Linear regression coefficient

Predictors	Coefficient
Intercept	-48.9543
X1	0.254265
X2	0.25535
X3	0.253065
X4	8.71E-08
X5	0.248216
X6	0.002462
X7	-3.85E-09

Volume and IHSG return have very small coefficient values. Because the previous section identified high multicollinearity among High, Low, Open, and SMA5, their coefficients cannot be interpreted as the independent effect of each predictor on the next-day TLKM closing price. However, for forecasting purposes, the results show that LR demonstrates the best predictive performance compared with the other models, achieving an RMSE of 81.38, a MAPE of 1.37%, and an R^2 of 0.928 on the

unseen test set.

Next, the RF model is interpreted using the feature importance shown in Figure 6. High and Low are the most important predictors, with importance values of 0.5950 and 0.3746, respectively. Together, these two predictors account for approximately 96.96% of the total feature importance, whereas the remaining predictors contribute much less to the RF prediction. However, these results should be interpreted cautiously because the price-related predictors exhibit substantial multicollinearity. High and Low collectively dominate the model's predictive power, but their individual importance values remain inherently coupled.

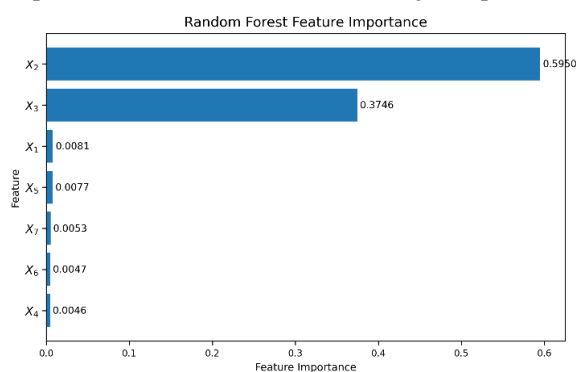


Figure 6. Random forest feature importance

3.6 Practical Implication

Based on the previous discussion, our findings provide a quantitative reference for forecasting the next-day TLKM closing price, with High and Low providing important predictive information. The forecasting model can be considered an alternative methodological tool for investors and researchers when evaluating short term stock price movements.

This study has several limitations. First, it focuses only on TLKM stock. Second, it is limited by the specific dataset period and the selected predictors. As a result, these findings cannot be directly generalized to other stocks or market periods. Future research could extend this study by incorporating additional market, technical and fundamental variables. In addition, future work could explore other machine learning models and longer forecasting horizons.

CONCLUSION

This study employs LR, RF and BPNN models to forecast the next day TLKM closing price. The models use seven predictors, namely

High, Low, Open, Volume, SMA₅, RSI₁₄, and IHSIG return. To determine suitable hyperparameter configurations, we applied RandomizedSearchCV with 50 randomly selected configurations to RF and a random search with 10 randomly selected configurations to BPNN. Both searches were conducted using time-series cross-validation to preserve the chronological order of the data.

The results show that LR has the best predictive performance compared with the RF and BPNN models, achieving an MAE of 57.67, an RMSE of 81.38, a MAPE of 1.37%, and an R² of 0.928 on unseen data. This indicates that, for the TLKM dataset and the selected period, LR provides a strong and relatively consistent performance. Furthermore, greater model complexity does not necessarily lead to better forecasting performance.

DECLARATIONS

Author Contributions: The first author: Conceptualization, methodology, writing, review and editing. The second author: Data curation, investigation, visualization and original draft writing. The third author: Conceptualization, original draft writing. The fourth author: Conceptualization, review and editing. All authors reviewed and approved the final manuscript.

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

Generative Artificial Intelligence Use Statement: No Generative AI or Large Language Model was used to generate the scientific content, research data, analysis, interpretation, or conclusions of this manuscript. Generative AI was used only for language checking and writing assistance. The authors remain fully responsible for the final content.

Ethics Approval and Informed Consent: Not applicable. This study did not involve human participants or personal data.

Data Availability: The data used in this research were obtained from publicly available Yahoo Finance data and are available from the corresponding author upon reasonable request.

REFERENCES

- [1] Z. Bodie, A. Kane, and A. J. Marcus, *Investments*, 13th ed. New York, NY, USA: McGraw Hill, 2024. ISBN: 978-1-266-08596-3.
- [2] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Systems with Applications*, vol. 197, Art. no. 116659, 2022, doi: 10.1016/j.eswa.2022.116659.
- [3] I. Himawan, O. Nurdian, and G. Dwilestari, "Prediksi harga saham dengan algoritma regresi linier dengan RapidMiner," *JURSIMA*, vol. 10, no. 3, 2022, doi: 10.47024/js.v10i3.475.
- [4] D. Ambarita, R. G. Marpaung, and J. P. Nainggolan, "Analysis of closing price forecasts for PT Bank Mandiri (Persero) Tbk (BMRI) stock using the multiple linear regression method based on OHLCV data," *Journal of Management Analytical and Solution*, vol. 6, no. 2, pp. 85–96, 2026, doi: 10.32734/jomas.v6i2.25006.
- [5] M. Hasbi, M. Jamil, M. H. Akbar, and Sharipuddin, "Analisis dan prediksi harga saham menggunakan metode Random Forest," *Jurnal Ilmiah Media Sisfo*, 2026.
- [6] Y. H. Parulian, B. Subartini, and Riaman, "Analisis pengaruh IHSG, inflasi, BI Rate dan nilai tukar Rupiah terhadap fluktuasi harga saham bank di Indonesia menggunakan metode regresi linear berganda," *Jurnal Matematika Integratif*, vol. 9, no. 1, pp. 19–28, 2013, doi: 10.24198/jmi.v9.n1.10195.19-28.
- [7] E. Fitri and D. Riana, "Analisa perbandingan model prediction dalam prediksi harga saham menggunakan metode Linear Regression, Random Forest Regression dan Multilayer Perceptron," *METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi*, vol. 6, no. 1, pp. 69–78, 2022, doi: 10.46880/jmika.Vol6No1.pp69-78.
- [8] C. Fieberg, D. Metko, T. Poddig, and T. Loy, "Machine learning techniques for cross-sectional equity returns' prediction," *OR Spectrum*, vol. 45, no. 1, pp. 289–323, 2023, doi: 10.1007/s00291-022-00693-w.
- [9] S. Suyatna, A. J. Satrio, S. P. Pratama, M. Indarto, F. Syarif, and F. Amsury, "Analisis perbandingan kinerja model machine learning dalam prediksi harga saham TLKM," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 4, pp. 829–837, 2025, doi: 10.31004/riggs.v4i4.3476.
- [10] S. Lumbantoruan, A. Duha, T. P. Astuti, A. J. Wahidin, B. Rahmatullah, and I. Kurniawati, "Prediksi harga saham PT Telkom Indonesia Tbk (TLKM) menggunakan Linear Regression, Random Forest, dan XGBoost," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 4, pp. 3429–3438, 2025, doi: 10.31004/riggs.v4i4.4158.
- [11] M. T. Pawitra, L. Abdurrahman, and H. Fakhurroja, "FinBERT-based sentiment integration in hybrid CNN-BiLSTM models for stock price forecasting," *Jurnal Teknik Informatika*, vol. 19, no. 1, pp. 73–84, 2026, doi: 10.15408/jti.v19i1.49466.
- [12] A. Q. Md, S. Kapoor, C. J. Chris, A. K. Sivaraman, K. F. Tee, H. Sabireen, and N. Janakiraman, "Novel optimization approach for stock price forecasting using multi-layered sequential LSTM," *Applied Soft Computing*, vol. 134, p. 109830, 2023, doi: 10.1016/j.asoc.2022.109830.
- [13] C. Li and G. Qian, "Stock price prediction using a frequency decomposition based GRU Transformer neural network," *Applied Sciences*, vol. 13, no. 1, p. 222, 2023, doi: 10.3390/app13010222.
- [14] T. Li, Z. Liu, Y. Shen, X. Wang, H. Chen, and S. Huang, "MASTER: Market-Guided Stock Transformer for Stock Price Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, pp. 162–170, 2024, doi: 10.1609/aaai.v38i1.27767.
- [15] J. V. R. Ferro, R. J. R. Dos Santos, E. B. Costa, and J. R. S. Brito, "Machine learning techniques via ensemble approaches in stock exchange index prediction: Systematic review and bibliometric analysis," *Applied Soft Computing*, vol. 167, p. 112359, 2024, doi: 10.1016/j.asoc.2024.112359.
- [16] R. Chopra, G. D. Sharma, and V. Pereira, "Identifying bulls and bears? A bibliometric review of applying artificial

- intelligence innovations for stock market prediction,” *Technovation*, vol. 135, p. 103067, 2024, doi: 10.1016/j.technovation.2024.103067.
- [17] C. Bergmeir, R. J. Hyndman, and B. Koo, “A note on the validity of cross-validation for evaluating autoregressive time series prediction,” *Computational Statistics & Data Analysis*, vol. 120, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.
- [18] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [19] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*. Cham, Switzerland: Springer, 2023.
- [20] C. Lewis, *Industrial and Business Forecasting Methods: A Radical Guide to Exponential Smoothing and Curve Fitting*, vol. 56(2), London: Butterworth Scientific, 1982.