SENTIMETER (Seminar Nasional Teknologi Informasi, Mekatronika dan Ilmu Komputer)
Universitas Nusa Putra

Comparison of Random Forest and LSTM Methods for Predicting Google Class C Stock Prices

Muhamad Fahraz Firdaus[1, a)], Malik Pajar Anugrah[2, b)], Najla Ghaida Fauziyah[3, c)] , Sri Wahyunil Khotimah[4, d)]

[1)] Fakultas Teknik Komputer dan Desain, Nusa Putra University
Sukabumi 43152, Indonesia

e-mail: muhamad.fahraz_ti22@nusaputra.ac.id[1)] , malik.pajar_ti22@nusaputra.ac.id[2)] ,
najla.ghaida_ti22@nusaputra.ac.id[3)] , sri.wahyunil_ti22@nusaputra.ac.id[4)]

* Korespondensi: e-mail: muhamad.fahraz_ti22@nusaputra.ac.id

ABSTRAK

Prediksi harga saham merupakan salah satu tantangan utama dalam analisis keuangan, terutama di pasar yang sangat dinamis dan tidak terduga. Penelitian ini bertujuan untuk membandingkan efektivitas dua metode pembelajaran mesin, yaitu Random Forest (RF) dan Long Short-Term Memory (LSTM), dalam memprediksi harga saham Google kelas C. Dengan menggunakan data historis harga saham dari 1 Januari 2020 hingga 30 Oktober 2024, penelitian ini menerapkan tahapan pengolahan data yang meliputi pengumpulan, pembersihan, dan transformasi data untuk memastikan kualitas dataset. Selain itu, tahapan feature engineering dilakukan untuk menambahkan fitur-fitur penting yang dapat meningkatkan akurasi prediksi. Data dibagi menjadi data pelatihan dan pengujian dengan proporsi 70:30 untuk menguji kemampuan generalisasi model. Model LSTM mencatat kinerja unggul dengan RMSE sebesar 3,02, MAE sebesar 2,24, MAPE sebesar 1,32%, dan R² sebesar 0,95, menunjukkan kemampuannya menangkap pola kompleks pada data sekuensial. Sebaliknya, Random Forest memiliki RMSE sebesar 9,34, MAE sebesar 7,37, MAPE sebesar 5,45%, dan R² sebesar 0,14, yang mengindikasikan keterbatasannya dalam menjelaskan variabilitas data. Ini menunjukan bahwa LSTM lebih sesuai untuk prediksi data deret waktu yang kompleks, sedangkan Random Forest lebih relevan untuk situasi yang membutuhkan interpretabilitas dan kecepatan komputasi. Pemilihan metode prediksi harus mempertimbangkan karakteristik data dan kebutuhan analisis.

Kata Kunci*: Prediksi Harga Saham, Random Forest, LSTM, Rekayasa Fitur, Analisis Keuangan*

ABSTRACT

Stock price prediction is one of the main challenges in financial analysis, especially in highly dynamic and unpredictable markets. This study aims to compare the effectiveness of two machine learning methods, namely Random Forest (RF) and Long Short-Term Memory (LSTM), in predicting the stock price of Google Class C shares. Using historical stock price data from January 1, 2020, to October 30, 2024, this research applies data processing stages, including data collection, cleaning, and transformation, to ensure dataset quality. Additionally, feature engineering is conducted to add important features that can enhance prediction accuracy. The data is split into training and testing sets with a 70:30 ratio to test the model's generalization capability. The LSTM model recorded superior performance with an RMSE of 3.02, MAE of 2.24, MAPE of 1.32%, and R² of 0.95, demonstrating its ability to capture complex patterns in sequential data. In contrast, Random Forest had an RMSE of 9.34, MAE of 7.37, MAPE of 5.45%, and R² of 0.14, indicating its limitations in explaining data variability. This suggests that LSTM is more suitable for predicting complex time series data, while Random Forest is more relevant for situations that require interpretability and computational speed. The choice of prediction method should consider the characteristics of the data and the needs of the analysis.

Keywords*: Stock Price Prediction, Random Forest, LSTM, Feature Engineering, Financial Analysis*

I. INTRODUCTION

Many scholars and professionals in the finance industry are interested in the topic of stock price prediction. Since the stock market's founding, financial scientists and sociologists from all over the world have regarded the market's growth rate as one of the most important markers for assessing the wealth and degree of development of a certain nation or area. As a result, research on the stock market is extremely useful [1]. Because of the stock market's notoriously erratic and non-linear behavior, predicting stock prices is an extremely difficult undertaking. The market's speculative activity, in which traders typically exploit brief price swings to make money, can be intensified by frequent stock price movements. In addition to raising stock market risk, this speculative activity may also cause increased volatility and instability, which would leave other investors in the dark [1].

One of the digital giants in the globe, Google Inc., is one business that has attracted a lot of interest in the stock market. Many investors now find Google's stock to be an appealing investment option due to its ongoing innovation and domination in a number of industries, including cloud services, digital advertising, and internet search. The stock of Google is regarded as a significant element in the current financial markets. Predicting future stock price changes is a topic of great interest, despite the fact that it is thought to be both difficult and complicated. In the hopes that future events can be predicted from historical data, researchers, the business community, and interested parties work to forecast stock prices [2].

Numerous deep learning models and machine learning algorithms have been created recently to improve the accuracy of stock price forecasting. Techniques like Long Short-Term Memory (LSTM) and Random Forest (RF) have shown promise in identifying intricate patterns in stock market data. To increase accuracy and lower the chance of overfitting, RF builds several decision trees and combines their output. Furthermore, RF is extremely versatile and efficient in a wide range of financial applications because to its ability to manage hundreds of input variables [3]. However, LSTM, a deep neural network type, is a desirable option for time series analysis, including stock price prediction, since it is made to handle sequence and long-term dependency problems in data. When processing sequential data, LSTM is quite effective. It can be further tuned to increase prediction accuracy in a variety of application domains [4].

Given this context, the study's objective is to evaluate how well the SVM and LSTM approaches forecast Google's stock values. It is anticipated that a better model with trustworthy and accurate predictions would be found through further investigation.

Based on the background provided, the research problem formulation can be outlined as follows:

1. How can the effectiveness of machine learning methods, specifically Random Forest (RF) and Long Short-Term Memory (LSTM), be compared in predicting stock prices for Google Class C shares?
2. What is the impact of using historical stock price data and feature engineering on the accuracy and reliability of stock price prediction models?
3. How do RF and LSTM handle the challenges of non-linear, unstable, and volatile stock market data differently, and which model provides better prediction accuracy for Google’s stock prices?
4. Can a comprehensive comparison of RF and LSTM models identify a superior model that offers accurate and reliable predictions for application in financial analysis and stock market forecasting?

II. LITERATURE REVIEW

Six deep learning models were examined in earlier studies to forecast stock prices in the S&P 500 market, which is renowned for its non-linearity, dynamism, and volatility [5]. This work brought attention to the difficulties in forecasting stock prices, a popular subject among technology and finance academics. Long Short-Term Memory (LSTM), Autoencoder (AE), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Multilayer Perceptron (MLP), and Gated Recurrent Unit (GRU) were the six models that were tested. The closing prices of 20 businesses during a seven-year span, from January 2015 to August 2022, made up the dataset. The purpose of the study was to investigate how well deep learning models can extract intricate patterns from unstructured data and to offer suggestions for how they might be used in financial markets. An additional study compared the performance of the Random Forest and Long Short-Term Memory (LSTM) models in predicting Tesla's stock prices, highlighting the significance of stock price prediction in the intricate and international financial market. Experimental results demonstrated that the RNN model performed best in predicting stock prices, with lower evaluation metrics such as Mean

Absolute Error (MAE) and Root Mean Squared Error (RMSE) compared to other models.

This research significantly advances our understanding of deep learning models in the stock market and opens opportunities for future research in optimizing stock price prediction techniques [6]. Time series analysis theory, a crucial technique for forecasting stock values from historical data, was used in this study. An ensemble-based technique called Random Forest combines several decision trees to increase accuracy. An artificial neural network called LSTM, on the other hand, is made to handle non-linear correlations and long-term dependencies in time series data.

Data collection using a publicly available dataset on Tesla stock prices from Kaggle, model evaluation using metrics like Mean Square Error (MSE) and R-Square (R^2), and problem identification pertaining to the requirement for precise stock price predictions were all included in the research framework. Because LSTM can capture long-term dependencies and non-linear interactions in time series data, the study predicted that it would have a higher prediction accuracy than Random Forest.

With an R2 of 0.943 and an MSE of 684,407, LSTM greatly beat Random Forest, according to the results. After feature lag optimization, Random Forest produced an R2 of 0.518 and an MSE of 706,409. The significance of choosing models that complement the distinct features of financial data is emphasized by this study, which also provides opportunities for additional research in predictive model optimization. As a result, the study offers a more accurate option for forecasting Tesla's stock values and makes substantial advances to our understanding of the use of predictive models in the financial sector.

III. Research Methodology

A. Research Method

The research approach used in this study adheres to the Cross-Industry Standard Process for Data Mining (CRISP-DM) paradigm, which is one of the standard methodologies in data mining research. Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment are the six phases of this model [7]. The process in this study, however, begins with the Data Understanding stage and continues through the Evaluation stage. The phases of the suggested research approach are as follows:

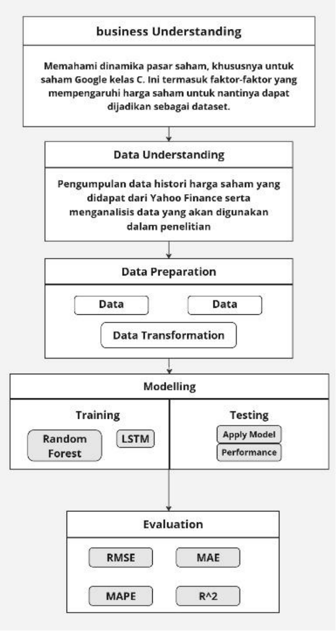
business Understanding
Memahami dinamika pasar saham, khususnya untuk saham Google kelas C. Ini termasuk faktor-faktor yang mempengaruhi harga saham untuk nantinya dapat dijadikan sebagai dataset.
Data Understanding
Pengumpulan data histori harga saham yang didapat dari Yahoo Finance serta menganalisis data yang akan digunakan dalam penelitian
Data Preparation
Data
Data
Data Transformation
Modelling
Training
Random Forest
LSTM
Testing
Apply Model
Performance
Evaluation
RMSE
MAE
MAPE
R^2

Figure 1. Research Methodology

B. Dataset Processing Method

In this study, data processing is essential to ensuring that the data used fits the prediction model's requirements. Model accuracy will increase and the probability of errors will decrease with clean and organized data. These are the procedures for handling the dataset:

1) Data Collection

Google Class C stock price historical time series data, obtained from the Yahoo Finance platform using the Python yahoo-finance module, served as the study's source of data. Key characteristics of the data, which includes open price, closing price, high price, low price, and trading volume, are available from January 1, 2020, to December 31, 2024. A wider range of patterns in stock price trends is intended to be provided by using historical data over an adequate amount of time.

2) Data Cleaning

The researcher finds and fixes any missing or inaccurate data in the dataset at this point. If not managed appropriately, missing data might interfere with analysis and affect model performance. Several techniques are employed, such as eliminating rows with missing data or imputing missing values (for instance, by using the mean or median). Changes or deletions are made if there are any outliers or anomalies.

3) Data Transformation

Following cleaning, the data is converted to facilitate processing by the prediction model. Normalization or standardization of the data is part of this transformation, especially for data-scale-sensitive models like LSTM. By scaling the data to a predetermined range, such 0 to 1, normalization makes it simpler and more consistent for the model to process the input. This step is essential to make sure that the weight assigned to each feature in the prediction model is unaffected by the scale between variables.

4) Feature Engineering

In order to increase prediction accuracy, new features that are deemed significant are added during the feature engineering process. The information includes technical elements including relative strength index (RSI), moving averages, and other indicators frequently used in stock research. Particularly for models that use time series data, such as LSTM, these new features aid the model in comprehending patterns of stock price movement.

5) Data Splitting

Training data and testing data are the two categories into which the processed data is separated. The model is usually trained using around 70% of the data, with the remaining 30% going toward testing. This division's goal is to test the model's generalization and prediction skills on fresh data by exposing it to data it has never seen before.

IV. Result and Discussion

The dataset is fed into the Random Forest and LSTM models for stock price prediction, depending on the approach employed. Seventy percent of the dataset is used for the training test model, while the remaining thirty percent is used for the testing model. Figure 2 displays the LSTM model's running visual result.

A. *LSTM Result*

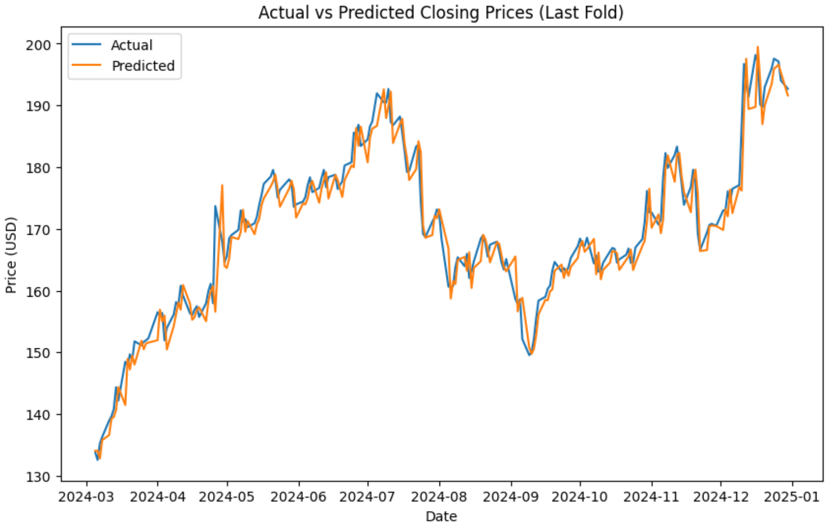

Figure 2. LSTM Model Graphic

The picture shows the difference between an asset's actual and anticipated closing prices between March 2024 and January 2025. The orange line represents the prices that a machine learning algorithm anticipated, while the blue line represents the actual closing prices. This plot graphically illustrates how successfully the model generated changes in the asset's price. The forecasts are good if the lines are fairly near together; if they diverge significantly, the predictions are not so good.

The first layer of the LSTM model we developed for this study contained 50 memory units that generated sequences, while the second layer was set up to only return the most recent output. The model was designed to handle input sequences with 60 time steps using a single feature. The Adam optimizer was utilized for efficient training, while Mean Squared Error was employed as the loss function to measure prediction accuracy. Over 100 epochs, the model was trained with a batch size of 32. With a Root Mean Squared Error (RMSE) of roughly 3.02, which indicates low average prediction error, the assessment metrics showed excellent performance. The mean absolute error (MAE), which measures the average size of prediction mistakes, was 2.24. Additionally, compared to the real data, the Mean Absolute Percentage Error (MAPE) was 1.32%, which was exceptionally low and showed good accuracy. The model's capacity to recognize complex patterns in the sequential data and its promise for practical time series forecasting applications were demonstrated by its R2 value of 0.95, which allowed it to explain about 95% of the variance in the target variable.

B. Random Forest Result

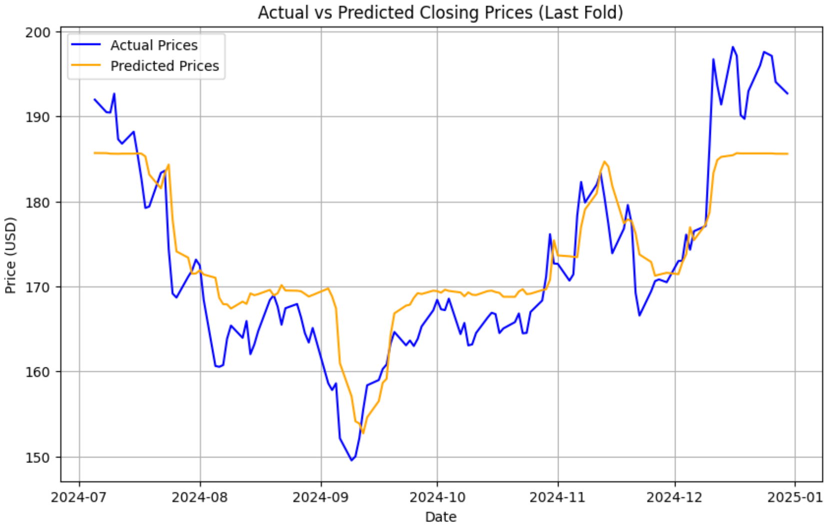

Figure 3. Random Forest Model Graphic

This graph illustrates the discrepancy between the actual and anticipated closing values of an asset between July and January of 2025. The actual closing prices are shown by the blue line, which represents the asset's price changes over time. The orange line indicates the closing prices that a machine learning algorithm anticipates. A visual comparison of the two lines can be used to assess the model's accuracy. The more closely the lines overlap, the more accurate the model's predictions are. On the other side, significant line divergence indicates that the model's predictions are less reliable.

the Random Forest model set up with particular hyperparameters; 200 estimators, a maximum depth of 20, a minimum samples split of 2, a minimum samples leaf of 2, and no upper restriction on maximum features. These parameters were chosen to achieve a compromise between model complexity and performance, allowing the ensemble approach to effectively find patterns in the data. The model's performance was evaluated using a variety of measures, and the average Root Mean Squared Error (RMSE), which displays the average discrepancy between predictions and actual values, was 9.34. The Mean Absolute Error (MAE), which averaged 7.37, represented the average magnitude of errors without accounting for their direction. Moreover, the Mean Absolute Percentage Error (MAPE), calculated at 5.45%, demonstrated the accuracy in respect to the real numbers. There is still significant room for improvement in identifying the underlying links in the data, even though the Random Forest model had some predictive power. The target variable's variation was only partially explained by the model, as seen by its average R2 score of 0.14.

C. Table Comparasion Between LSTM and Random Forest

Algorithm	**Parameter**	**RMSE**	**MAE**	**MAPE**	**R^2**	**Computation Time**
LSTM	Units: 50, Return Sequences: True (first layer), False (second layer), Input Shape: (60, 1), Optimizer: Adam, Loss Function: Mean Squared Error Batch Size: 32 Epochs: 100	3.0239	2.2440	1.3160	0.9473	29 sec
Random Forest	n_estimators: 200, min_samples_split: 2, min_samples_leaf: 2, max_features: None, max_depth: 20	9.3360	7.3655	5.4546	0.136	39 sec

V. Conclusion

Regarding their ability to forecast time series, the Random Forest and LSTM models' results clearly demonstrate their benefits and drawbacks. The LSTM model achieved impressive evaluation metrics, including a Root Mean Squared Error (RMSE) of approximately 3.02, a Mean Absolute Error (MAE) of approximately 2.24, a Mean Absolute Percentage Error (MAPE) of 1.32%, and an R2 score of 0.95. These metrics demonstrated that the model was able to capture complex patterns and explain a significant portion of the variance in the target variable. With an average RMSE of 9.34 and an R2 score of 0.14, the Random Forest model performed admirably but fell short of LSTM in identifying the underlying relationships in the data. These findings are in line with earlier studies that demonstrate that deep learning models, such as LSTM, often outperform more traditional machine learning approaches, such as Random Forest, for sequential data problems. However, Random Forest has demonstrated advantages in terms of minimizing bias and adapting to variances, which may make it superior in some circumstances when interpretability and speed are crucial factors. Even though LSTM appears to be superior for complex time series forecasting tasks, Random Forest may still be helpful depending on the specific context and analysis requirements. These findings suggest that for time-series forecasting problems including complex sequential data, LSTM models would be a better fit in this scenario than Random Forest.

VI. Bibliography

[1] Z. Luo, “The Prediction of Google Stock Closing Price Based on Linear Regression Model and Random Forest Model,” no. Icdse, hal. 229–233, 2024, doi: 10.5220/0012805600004547.

[2] Y. Huang, “Research on the Google Stock Price Prediction Based on SVR, Random Forest, and KNN Models,” *Highlights Business, Econ. Manag.*, vol. 24, hal. 1054–1058, 2024, doi: 10.54097/n8hxqx19.

[3] S. Ji, “Predict stock market price by applying ANN, SVM and Random Forest,” *SHS Web Conf.*, vol. 196, hal. 02005, 2024, doi: 10.1051/shsconf/202419602005.

[4] S. M. Al-Selwi *et al.*, “RNN-LSTM: From applications to modeling techniques and beyond—Systematic review,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 5, hal. 102068, 2024, doi: 10.1016/j.jksuci.2024.102068.

[5] V.-T. Duong, D.-T.-A. Nguyen, T.-T.-H. Pham, V.-H. Nguyen, dan V.-Q. A. Le, “Comparative Study of Deep Learning Models for Predicting Stock Prices,” *Proc. Seventh Int. Conf. Res. Intell. Comput. Eng.*, vol. 33, hal. 103–108, 2023, doi: 10.15439/2022r02.

[6] X. Wang, “Stock Price Prediction: A Comparative Study of Random Forest and LSTM Models,” *Highlights Sci. Eng. Technol.*, vol. 107, hal. 117–123, 2024, doi: 10.54097/70a8b947.

[7] R. Wirth dan J. Hipp, “CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39,” *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, hal. 29–39, 2000, [Daring]. Tersedia pada: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining