The Use of Optimized Random Forest in Analyzing the Prediction of Stock Trends in Indonesia

Anak Agung Ayu Diva Shanty Darmawan[1, a)], Kie Van Ivanky Saputra[1, b)], and Ferry Vincenttius Ferdinand[1, c)]

[1]*Faculty of Science and Technology*
Pelita Harapan University
MH Thamrin Boulevard 1100, Klp. Dua, Kec. Klp. Dua, Kota Tangerang, Banten 15811.

[a)]01112190018@student.uph.edu
[b)] kie.saputra@uph.edu
[c)] ferry.vincenttius@uph.edu

Abstract. Stocks are ownership symbols of an individual or an institution, which are long-term financial instruments and can be traded. This research employs the optimized random forest as the primary method to assist in predicting stock trends in the energy sector. The stocks used are ADRO.JK, INDY.JK, PTBA.JK, TOBA.JK, and UNTR.JK. The first step involves identifying the technical indicators that will be used as variables to aid in predicting stock trends, using adjusted closing price of the stocks. The model will be constructed using two sets of data: stock-based and year-based data. There are four stages in the model development: random forest model, optimized random forest model, random forest model with feature importance, and optimized random forest with feature importance. The model will be evaluated based on accuracy, F1 score, and AUC value. The results obtained from the optimized random forest method will be compared with those from the random forest method. This research demonstrates that the optimized random forest is a more accurate method for predicting stock trends. Furthermore, the predictions from the best-performing model will be used in a stock trading simulation.

Keywords: Random Forest, Optimized Random Forest, Classification, Stock, Technical Indicator, Multiclass.

I. INTRODUCTION

The term "capital market" refers to a market where long-term financial instruments can be traded. There are various types of capital market instruments, one of which is stocks [1]. Stocks are a form of securities that represent ownership by individuals or institutions in a company [2]. There are many stock options available to investors, and typically, each investor has their own preferences when selecting stocks. After conducting a selection, the question arises about the right time for an investor to buy and sell the chosen stocks. One of the methods for determining this is by analyzing stock trends.

A stock trend is a condition in which there is a dominating movement in the stock price chart, and this can occur continuously [3]. However, in understanding stock trends, it is not sufficient to simply observe them to predict the right time to buy or sell selected stocks. Therefore, knowledge of the types of stock trend patterns and strategies suitable for investors is required to make an informed decision. However, analyzing and predicting stock trends with this knowledge can still be challenging. This is because the stock market is dynamic and complex, with a constant influx of new information about stocks every day [4].

Many methods have been developed to assist investors in predicting stock trends. Data mining and machine learning have often been applied to build effective models, such as support vector machine and neural networks [4]. With the multitude of methods available, the issue to be investigated is which method should be applied to produce a model that can achieve the most accurate prediction results. This research will propose the optimized random forest method to build an accurate model for predicting stock trends in Indonesia.

Random forest is a method that combines the performance of different decision tree algorithms to classify or predict the value of a variable [5]. Optimized random forest is an extension of the random forest method with additional steps such as parameter optimization and feature extraction to achieve higher accuracy results. In this research, the optimized random forest method will be used to analyze daily stock data from five energy companies listed on the Indonesia Stock Exchange (BEI). Additionally, a comparison will be made between the model generated by optimized random forest and the model from the traditional random forest to determine which model has a higher level of accuracy.

II. BACKGROUND

A. Technical Indicators

Predicting stock trends requires an analytical method that can interpret stock price movements over a specific period, commonly known as technical analysis [6]. This method relies on parameters obtained from stock data, referred to as technical indicators [4]. Technical indicators are tools commonly used to evaluate the short-term dynamics of stock prices. These tools also prove effective in their application to stocks with medium and long-term perspectives [4]. In this research, technical indicators serve as independent variables. The following are the technical indicators used.

Table 1. Types of Technical Indicators

Technical Indicators	Full Name of Index	Technical Indicators	Full Name of Index
EMA	Exponential moving average	BOLL	Bollinger bonds
MACD	Moving average convergence divergence	OBV	On balance volume
MTM	Momentum index	SMA	Simple moving average
RSI	Relative strength index	TSI	True strength index
ATR	Average true range	Coppock	Coppock curve
ADOSC	A/D oscillator	SO	Stochastic oscillator
CCI	Commodity channel index	Mass index	Mass index
MFI	Money flow indicator	EMV	Ease of movement value
ULTOSC	Ultimate oscillator	VI	Vortex indicator
Donchian	Donchian channel	Chaikin	Chaikin oscillator

B. Random Search

Random search is an algorithm that utilizes various types of probabilities to aid in decision-making **[7]**. This algorithm serves as an alternative to grid search. In this research, random search will be employed to optimize parameters to achieve a high level of accuracy in the constructed model. The search range for random search in this study includes the number of trees (*n_estimators*), maximum tree depth (*max_depth*), the maximum number of features for splitting (*max_features*), the minimum number of samples required to split an internal node (*min_samples_split*), and the minimum number of samples required to be in a leaf node (*min_samples_leaf*).

C. Random Search

Random Forest (RF) is a machine learning technique that combines various decision tree algorithms to classify or predict the values of a variable [5]. RF is a classic ensemble learning model and is more accurate than a single decision tree [4]. The bagging procedure is what sets RF apart from a standalone decision tree. Bootstrap aggregating (bagging) is a technique used in machine learning to split a dataset into several equally sized random subsets through a process of sampling with replacement (bootstrap).

Optimized Random Forest is a refined version of the random forest method that has been optimized to achieve higher model accuracy. The optimization involves tuning the model parameters, typically performed using a random search method before being applied to the model. The model with parameters that yield the highest accuracy is selected for testing on a separate test set using evaluation tests.

D. Literature Review

1. Research by Lili Yin, Benling Li, and Rubo Zhang titled 'Research on Stock Trend Prediction Method Based on Optimized Random Forest'. In this study, the Optimized Random Forest method was tested and compared with the normal Random Forest method and the Light Gradient Boosting method. The research found that the optimized random forest method had the highest accuracy rate for each stock data compared to the other two methods [4].
2. Research by Suryoday Basak, Saibal Kar, Snehansu Saha, Luckyson Kaidem, and Sudeepa Roy Dey titled 'Predicting the Direction of Stock Market Prices Using Tree-Based Classifiers'. This study compared several tree-based classifiers. It was found that the random forest method and the XGBoost method had higher accuracy rates and better performance compared to other methods [8].
3. Research by Michael Ballings, Dirk Van den Poel, Nathalie Hespeels, and Ruben Gryp titled 'Evaluating Multiple Classifiers for Stock Price Direction Prediction'. The research compared Ensemble classification methods with Single classification methods. Through this, it was found that the random forest method had the highest accuracy rate in predicting the forecasted direction of stock prices [9].

III. RESEARCH METHOD

A. Data

The data used in this research consists of daily energy stock data collected from January 1, 2016, to December 31, 2021. The daily stock data includes opening price, closing price, lowest price, adjusted close price, and stock trading volume. The stock data was obtained from the Yahoo Finance website. The selected stocks are energy stocks listed on the Indonesia Stock Exchange (BEI), namely ADRO.JK, INDY.JK, PTBA.JK, TOBA.JK, and UNTR.JK. The total number of collected data points is 1626, with six columns for each of the five selected stocks. Below is a snippet of the initial ten data points for the ADRO.JK stock that has been collected.

Table 2. ADRO.JK Stock Data

Date	Open	High	Low	Close	Adjusted Close	Volume
02/01/2017	1695	1695	1695	1695	1065.46	0
03/01/2017	1690	1750	1685	1740	1093.74	64546200
04/01/2017	1730	1735	1690	1700	1068.60	45023900
05/01/2017	1695	1710	1660	1665	1046.60	34668500
06/01/2017	1660	1705	1655	1695	1065.46	24130100
09/01/2017	1700	1720	1685	1720	1081.17	25338200
10/01/2017	1695	1705	1680	1685	1059.17	19228600
11/01/2017	1720	1760	1700	1745	1096.89	54843500
12/01/2017	1750	1750	1715	1715	1078.03	22256200
13/01/2017	1730	1735	1685	1700	1068.60	27517400

In this research, the data to be used is divided into two types: stock-based data and year-based data. Stock data consists of daily data for each stock from January 1, 2017, to December 31, 2021. Year data is stock data collected based on the respective years. The following table is a summary table of each type of data along with their respective quantities.

Table 3. Research Data Summary

Data Type	Category	Data Total
Stock-Based	ADRO.JK	1262
	INDY.JK	1262
	PTBA.JK	1262
	TOBA.JK	1262
	UNTR.JK	1262
Year-Based	2017	1270
	2018	1305
	2019	1290
	2020	1210
	2021	1235

B. Feature Engineering

The stock data in question comprises the opening price, highest price, lowest price, closing price, adjusted close price, and stock volume. This data cannot be used as independent and dependent variables in this study. At this stage, the stock data will be utilized to identify independent and dependent variables.

The dependent variable used is close price prediction, utilizing the return from the closing price. Returns will be constrained within an upper limit of 2 and a lower limit of -2. If the return exceeds the upper limit, the resulting value is +1. If the return falls below the lower limit, the resulting value is -1. Otherwise, it results in a value of 0. The calculation of the dependent variable is as follows.

$$pred_1 = \begin{cases} +1, \ if \ \dfrac{C_2 - C_1}{C_1} \times 100 \geq 2 \\ 0, \ if -2 < \dfrac{C_2 - C_1}{C_1} \times 100 < 2 \\ -1, \ if \ \dfrac{C_2 - C_1}{C_1} \times 100 \leq -2 \end{cases}$$

$$\vdots$$

$$pred_{1261} = \begin{cases} +1, \ if \ \dfrac{C_{1262} - C_{1261}}{C_{1261}} \times 100 \geq 2 \\ 0, \ if -2 < \dfrac{C_{1262} - C_{1261}}{C_{1261}} \times 100 < 2 \\ -1, \ if \ \dfrac{C_{1262} - C_{1261}}{C_{1261}} \times 100 \leq -2 \end{cases}$$

After obtaining the dependent variable, the data will undergo a smoothing process. The data will be processed first using the exponential smoothing method. This is done to assist in identifying a trend. The type of exponential smoothing that will be used is simple exponential smoothing (SES). Following the smoothing results, the data will be used to find the independent variables for this study. The independent variables used are various technical indicators mentioned earlier.

The resulting SES data will be calculated, resulting in 26 technical indicators that will serve as independent variables for this study. The adjusted close price is the data used to calculate the independent variables, not the close price. The period used in this study is 15 periods, so at this stage, there are also some data that is deleted because it is used in the calculation to find the value of the technical indicators in the 15th period.

C. Data Preprocessing

In this stage, the dependent and independent variables of the data will be analyzed first. The analysis will be conducted by visually examining both variables using histogram and boxplot plots, and by observing the distribution of each variable. The results of this analysis will aid in handling missing values, non-existent values, and outliers. This is done to prevent errors in the analysis and to ensure that the final accuracy of the built model is not compromised.

D. Data Splitting

Before the data is split into a training set and a test set, the data will be grouped according to the year. Afterward, the data will be divided with a ratio of 70:30, where 70% will be the training set, and 30% will be the test set. The data split will be done randomly with the assistance of a Python program.

E. Model Building

1. Random Forest

   The construction of a random forest model typically involves creating a varying number of trees, commonly ranging from 10, 50, 100, up to 1000. The selected value for *min_sample_split*, which is the value used to split each node, is commonly chosen to be 3, 5, and 7. Each step of training and testing the model will be performed using a Python program.

2. Optimized Random Forest

   The construction of this model is not significantly different from building a random forest model. The differing behavior in this model lies in the parameter selection stage. There are several parameters that will be optimized with a larger range of parameters compared to the random forest model construction. In this stage, the parameters to be optimized include the number of trees, the maximum depth of the trees, the minimum number of samples required to split an internal node, and the minimum number of samples required to be in a leaf node. The parameters in building this model are not selected manually but will be chosen through a random search program in the Python application. Random search will assist in finding parameters that result in a model with higher accuracy. K-Fold Cross Validation will also be used in this random search so that the generated parameters have higher accuracy.

Table 4. Range of Random Search Parameters

Parameter	Parameter Name	Parameter Value
The number of trees in random forest	*n_estimators*	2, 4, 6, 8, 10, 12, 14, …, 1000
Maximum depth of trees	*max_depth*	2, 4, 6, 8, 10, …, 300
The maximum number of features for splitting	*max_features*	sqrt, log2, 0.25, 0.5, 0.75, 1, None
The minimum number of samples needed to split internal node	*min_samples_split*	2, 4, 6, 8, 10, …, 100
The minimum number of samples needed to be in leaf node	*min_samples_leaf*	2, 4, 6, 8, 10, …, 100

3. Feature Selection

   This is the stage where the variables that can be used as independent variables in model development will be determined. Several types of technical indicators will be applied as variables to assist in interpreting stock movements. Data will be run through these technical indicators, and they will be selected through feature importance analysis. The selected indicators are those that have passed a predetermined degree of importance. These selected indicators will become input features in the model to be constructed.

F. Model Testing and Evaluation

After both models are constructed, they will be tested using a test set. All results from both models will be compared with their actual values and recorded in a confusion matrix table. These outcomes will be evaluated using several metrics obtained from the confusion matrix and from the ROC curve. Accuracy, F1 Score, as well as AUC values, will be used as benchmarks in drawing conclusions.

IV. ANALYSIS AND DISCUSSION

A. Stock-Based Data Analysis

1. Random Forest

The first step taken with the clean stock data that has been rid of outliers is to split it into two parts: a training set and a test set, randomly using a 70:30 ratio. The results will be utilized in the next step, which involves constructing a random forest model using the training set with predefined parameters.

Table 5. Random Forest Parameter for Stock-Based Data

	n_estimators	**max_depth**	**max_features**	**min_samples_split**	**min_samples_leaf**
Default	200	20	None	6	4

Based on the model that has been built with specified parameters, the test data is used to evaluate the model through various evaluation tests that have been discussed. The confusion matrix table below is the result of the random forest model using TOBA.JK's data. Examining the results, a total of 291 samples were correctly predicted out of a total of 361 samples. It is evident from the results that the model often misclassifies samples that belong to class -1 or class +1 as class 0.

Table 6. TOBA.JK Random Forest Confusion Matrix

TOBA.JK				
		Predicted		
		-1	0	+1
Actual	-1	3	33	0
	0	1	288	0
	+1	1	35	0

Through the confusion matrix table, additional evaluation indices were obtained to assist in assessing the model's performance. Evaluation tests are used to test the model, and the following are the results. The classification results indicate that the model has not been able to learn the classification boundaries between classes well. The results show that only TOBA.JK stock has the highest accuracy among other stocks. The AUC value from the model evaluation results also indicates that the TOBA.JK stock model has relatively good ability to predict each class, as its value approaches 1.

Table 7. TOBA.JK Random Forest Model Evaluation

TOBA.JK				
Accuracy	Precision	Recall	F1-Score	AUC
0.8061	0.7083	0.8085	0.7275	0.8546

2. Optimized Random Forest

The results obtained from the random forest model are not very satisfactory. Therefore, to improve it, optimization of parameters is applied to search for parameters that can produce a model with a higher accuracy level. Parameter optimization is carried out with the help of random search. The following table shows the parameters generated by the random search used for each stock.

Table 8. Optimized Random Forest Parameters for Stock-Based Data

Stock	n_estimators	max_depth	max_features	min_samples_split	min_samples_leaf
ADRO.JK	400	26	log2	22	4
INDY.JK	650	286	None	14	2
PTBA.JK	902	66	sqrt	14	2
TOBA.JK	350	22	None	6	2
UNTR.JK	50	98	sqrt	10	2

The table below represents the confusion matrix results from the optimized random forest model with TOBA.JK's data. Examining the confusion matrix for TOBA.JK stock, it can be observed that there is an improvement in the number of correct classifications, especially in class +1, compared to the results of the random forest.

Table 9. TOBA.JK Optimized Random Forest Confusion Matrix

TOBA.JK				
		Predicted		
		-1	0	+1
Actual	-1	2	32	2
	0	3	286	0
	+1	0	33	3

The results from the previous confusion matrix table yield values for other evaluation metrics that can help assess the model's performance. The result table indicates that the TOBA.JK stock still has the highest accuracy among the stock-based data. Furthermore, it can be observed that stocks like TOBA.JK have experienced an improvement in other evaluation metrics. The increase in precision and recall values signifies that the model using the optimized random forest method performs better.

Table 10. TOBA.JK Optimized Random Forest Model Evaluation

TOBA.JK				
Accuracy	Precision	Recall	F1-Score	AUC
0.8061	0.7482	0.8065	0.7374	0.8546

3. Feature Selection

In this section, the model will be examined, and it will be rebuilt using independent variables that have the highest relative values with the model. Feature importance is utilized in this stage to enumerate these independent variables. Variables are selected based on the model used, so for each model, the independent variables may differ. A degree of importance of 0.04 is used as the threshold for variable selection. If a variable exceeds this threshold, it is chosen. The discussion will be divided into two based on the models, namely Random Forest with Feature Importance (RFFI) and Optimized Random Forest with Feature Importance (ORFFI).

a. Random Forest with Feature Importance (RFFI)

This model will be constructed based on the random forest model using the previously described stock-based data. Variables with a degree of importance exceeding 0.04 will be selected. The selected independent variables are chosen due to several influencing factors such as the length of the period used and the economic conditions in the market at that time.

Table 11. Independent Variables for Random Forest Model Stock-Based Data

Stock	Independent Variables
ADRO.JK	CO; 15-EMV; 1-EMV; ULTOSC; +VI; CC; MFI; -VI
INDY.JK	MI; 1-EMV; RSI; 15-EMV; +VI; CC; CCI; TSI; MACD; SO%k
PTBA.JK	1-EMV; MI; ATR; ULTOSC; +VI; 15-EMV; OBV
TOBA.JK	15-EMV; CCI; ATR; 1-EMV; RSI; +VI; ULTOSC; MFI; ADOSC; MACD
UNTR.JK	1-EMV; 15-EMV; MI; SO%k; OBV; +VI; CCI

The model will be evaluated using a test set through several evaluation tests. The table below shows the confusion matrix results of RFFI using stock-based data with stocks indicating the highest results. It can be seen from the table that there are a total of 290 samples correctly predicted out of 361 samples of TOBA.JK stocks.

Table 12. TOBA.JK Random Forest with Feature Importance Confusion Matrix

TOBA.JK				
		Predicted		
		-1	0	+1
Actual	-1	2	34	0
	0	1	288	0
	+1	0	36	0

The result of the confusion matrix can also yield evaluation indices that help assess the performance of the model. The test set is used to evaluate the model, and the table below shows the evaluation results for TOBA.JK stock.

Table 13. TOBA.JK Random Forest with Feature Importance Model Evaluation

TOBA.JK				
Accuracy	Precision	Recall	F1-Score	AUC
0.8033	0.7073	0.8065	0.7225	0.8525

b. Optimized Random Forest with Feature Importance (ORFFI)

The model will be constructed using an optimized random forest model obtained using stock market data. Variables with a degree of importance exceeding 0.04 will be selected. The selected independent variables are chosen based on factors that influence them, such as the length of the period used and the economic conditions in the market at that time.

Table 14. Independent Variables for Optimized Random Forest Stock-Based Data

Stock	**Independent Variables**
ADRO.JK	CO; +VI; 15-EMV; ULTOSC; CC; 1-EMV; MFI; -VI
INDY.JK	1-EMV; MI; 15-EMV; +VI; CCI; ULTOSC; RSI; CC; -VI; OBV; TSI; CO; MFI
PTBA.JK	+VI; MI; 1-EMV; ULTOSC; ATR; OBV; 15-EMV; CCI; CC; ADOSC
TOBA.JK	15-EMV; CCI; 1-EMV; ATR; EMA; +VI; RSI; ADOSC; ULTOSC; MI
UNTR.JK	1-EMV; 15-EMV; OBV; MI; SO%k; CO; ATR; MTM

Next, the model is evaluated using a test set with several evaluation tests. The following table shows the confusion matrix results of the optimized random forest model with feature importance using TOBA.JK's data. There are 291 samples that were correctly predicted out of a total of 361 samples.

Table 15. TOBA.JK Optimized Random Forest with Feature Importance Confusion Matrix

TOBA.JK				
		Predicted		
		-1	0	+1
Actual	-1	2	34	0
	0	1	288	0
	+1	1	34	1

The confusion matrix also yields values for evaluation indices that help assess the model's performance. The table below shows the results of this evaluation for TOBA.JK stock.

Table 16. TOBA.JK Optimized Random Forest with Feature Importance Model Evaluation

TOBA.JK				
Accuracy	Precision	Recall	F1-Score	AUC
0.8061	0.7980	0.8095	0.7275	0.8546

B. Year-Based Data Analysis

1. Random Forest

The first step after obtaining clean stock data without outliers is to split it into two parts: a training set and a test set, randomly using a 70:30 ratio. The results will be used in the next step, which involves building a random forest model using the training set with predetermined parameters.

Table 17. Random Forest Parameter for Year-Based Data

	n_estimators	max_depth	max_features	min_samples_split	min_samples_leaf
Default	200	20	None	6	4

Based on the built model, test data is used to assess the model, which is then evaluated. The following table shows the confusion matrix results of the model with data from the year. It can be seen from the data in 2019 that there are 259 samples predicted correctly out of a total of 370 samples.

Table 18. Year 2019 Random Forest Confusion Matrix

2019				
		Predicted		
		-1	0	+1
Actual	-1	3	45	0
	0	8	254	6
	+1	4	48	2

The confusion matrix results also yield evaluation index values that can help assess the model's performance. The test set is used for evaluation, and the 2019 results table has the highest accuracy. The AUC value is also above 0.7, indicating that the model has a good ability to correctly classify samples into their respective classes.

Table 19. Year 2019 Random Forest Model Evaluation

2019				
Accuracy	Precision	Recall	F1-Score	AUC
0.7000	0.5912	0.7017	0.6229	0.7750

2. Optimized Random Forest

The results obtained from the random forest will be further refined with different parameters to search for higher accuracy. Parameter optimization can be applied to find different parameters, and this is done with the help of random search. The following is a table of parameters generated by the random search used for each stock.

Table 20. Optimized Random Forest Parameters for Year-Based Data

Year	n_estimators	max_depth	max_features	min_samples_split	min_samples_leaf
2017	400	28	None	6	2
2018	400	26	None	18	2
2019	300	26	None	8	2
2020	900	72	0.75	26	2
2021	200	274	None	26	6

The test set will be used to evaluate the model with the described evaluation test. The following table shows the confusion matrix results of the optimized random forest model with year-based data. It can be seen that in the year 2019, 261 data points were correctly predicted out of a total of 370 samples. From the confusion matrix results, there is an increase in the random forest performance in both the +1 and -1 classes.

Table 21. Year 2019 Optimized Random Forest Confusion Matrix

2019				
		Predicted		
		-1	0	+1
Actual	-1	4	43	1
	0	10	253	5
	+1	6	44	4

The confusion matrix results also yield evaluation index values that can help assess the performance of the optimized random forest model with year data. The table below presents the evaluation results, showing an improvement in accuracy compared to the random forest model. The same trend is observed for precision and recall values, indicating an enhancement in the ability and precision of each model to classify samples into their respective classes.

Table 22. Year 2019 Optimized Random Forest Model Evaluation

2019				
Accuracy	Precision	Recall	F1-Score	AUC
0.7054	0.6203	0.7015	0.6343	0.7791

3. Feature Selection

In this section, the rebuilt model will be examined using independent variables that have the highest relative values with the model. Feature importance is utilized in this stage to enumerate these independent variables. Variables are selected based on the model used, so for each model, the independent variables may differ. A degree of importance of 0.04 is employed as the threshold for variable selection; if a variable surpasses this threshold, it is chosen. The discussion will be divided into two based on the models, namely Random Forest with Feature Importance (RFFI) and Optimized Random Forest with Feature Importance (ORFFI).

a. Random Forest with Feature Importance (RFFI)

This model will be built using a random forest model obtained using year-based data. Variables with a degree of importance exceeding 0.04 will be selected. The selected independent variables are chosen based on factors such as the length of the period used and the economic conditions at that time. The following are the selected independent variables from the random forest model with year-based data.

Table 23. Independent Variables for Random Forest with Feature Importance Year-Based Data

Year	Independent Variables
2017	MFI; CCI; +VI; ULTOSC; OBV; 1-EMV; RSI; EMA; ADOSC; CC; SO%k
2018	+VI; 15-EMV; OBV; MI; 1-EMV; ULTOSC; SO%k; ATR; CO
2019	MI; RSI; 1-EMV; SO%k; TSI; OBV; CC; ULTOSC; 15-EMV; ADOSC; MACD; CO; +VI; MTM
2020	-VI; OBV; 1-EMV; 15-EMV; ADOSC; MI; CC; EMA; Lower BB
2021	+VI; MI; 15-EMV; CC; CCI; OBV; ULTOSC; ATR

The test set will be used to evaluate the model with various evaluation tests. The table below shows the confusion matrix results from the 2019 data. It can be seen that there are a total of 262 samples correctly predicted out of a total of 370 samples.

Table 24. Year 2019 Random Forest with Feature Importance Confusion Matrix

2019				
		Predicted		
		-1	0	+1
Actual	-1	3	45	0
	0	7	258	3
	+1	2	51	1

The confusion matrix results can also yield evaluation indices that can assist in assessing the performance of a model. The following table presents the evaluation results of the model using data from the year 2019.

Table 25. Year 2019 Random Forest with Feature Importance Model Evaluation

2019				
Accuracy	Precision	Recall	F1-Score	AUC
0.7081	0.5977	0.7061	0.6185	0.7405

b. Optimized Random Forest with Feature Importance (ORFFI)

This model will be built using an optimized random forest model obtained from data for the year. Variables that have a degree of importance exceeding 0.04 will be selected. The selected independent variables are chosen due to several factors that influence them, such as the length of the period used and the economic conditions in the market at that time. The following table shows the selected independent variables from the optimized random forest model using year-based data.

Table 26. Independent Variables for Optimized Random Forest with Feature Importance Year-Based Data

Year	Independent Variables
2017	1-EMV; CCI; ULTOSC; MFI; MI; +VI; OBV; 15-EMV; EMA; CC; CO; -VI; RSI
2018	+VI; 15-EMV; OBV; MI; ULTOSC; 1-EMV; CCI; SO%k
2019	OBV; MI; 1-EMV; SO%k; RSI; CC; TSI; ADOSC; -VI; CO; ULTOSC; +VI
2020	OBV; ADOSC; -VI; 1-EMV; MI; 15-EMV; CCI; Lower BB; CC; ULTOSC
2021	+VI; MI; 15-EMV; CCI; OBV; CC; EMA; Upper BB; CO; ULTOSC

The test set will be used to evaluate the model with various performance tests. The following table shows the confusion matrix results of the optimized random forest model with feature importance from the 2019 data.

Table 27. Year 2019 Optimized Random Forest with Feature Importance Confusion Matrix

2019				
		Predicted		
		-1	0	+1
Actual	-1	4	43	1
	0	9	254	5
	+1	5	48	1

The confusion matrix also yields other evaluation metrics that can help assess the performance of the generated model. The test set is used for evaluation, and the table below represents the results of this evaluation based on the data from the year 2019.

Table 28. Year 2019 Optimized Random Forest with Feature Importance Model Evaluation

2019				
Accuracy	Precision	Recall	F1-Score	AUC
0.7000	0.5850	0.7014	0.6211	0.7750

C. Final Analysis

This section describes the analysis of all the data along with the previously explained models. The selected model is the one with the superior method for each dataset. The comparison for the best data and methods will be explained in this section. The following table summarizes all these results.

Table 29. Evaluation Summary from All Data

Data Type	Category	Model	Accuracy	F1-Score	AUC
Stock-Based	ADRO.JK	O-RF-FI	0.5961	0.4614	0.6971
	INDY.JK	O-RF-FI	0.5693	0.5109	0.6770
	PTBA.JK	RF-FI	0.6726	0.5868	0.7544
	TOBA.JK	O-RF	0.8061	0.7374	0.8546
	UNTR.JK	RF-FI	0.6494	0.5345	0.7478
Year-Based	2017	RF-FI	0.7190	0.6145	0.7892
	2018	O-RF	0.6207	0.5180	0.7155
	2019	O-RF	0.7054	0.6343	0.7791
	2020	RF-FI	0.5776	0.5201	0.6832
	2021	O-RF-FI	0.6803	0.5827	0.7602

The description of the table is about Random Forest (RF), Optimized Random Forest (O-RF), Random Forest with Feature Importance (RF-FI), and Optimized Random Forest with Feature Importance (O-RF-FI). In the table below, the best method is selected based on the number of model evaluations greater than the three evaluation indicators on a dataset. The table shows that the Optimized Random Forest method, with or without feature importance, outperforms the Random Forest method with feature importance on six datasets, which is superior to the latter on four datasets. When comparing the two datasets, the year-based data is considered superior to stock-based data because its average accuracy value is higher, namely 0.6606 compared to 0.6587 for stock-based data.

D. Trading Simulation

Stock simulation will be explained in this section. The built model will predict the dependent variable from the research that will serve as signals to buy or sell. Then, a stock trading simulation will be conducted using the predictions of the dependent variable over a one-month period. The Optimized Random Forest model with year 2019 data will be utilized to predict stock trends, with the prediction period starting from January 1, 2020, to February 29, 2020.

The algorithm will be implemented to predict the performance of five stocks over a one-month period. The rules applied in this simulation are as follows: if the prediction yields a value of +1, the action taken is to buy the stock. If the prediction yields a value of -1, the action taken is to sell the stock. Otherwise, the action taken is to stay inactive. The following table shows the predictions for TOBA.JK over a one-month period, assuming that 50 lots will be purchased for each buy action. TOBA.JK, compared to other stocks, exhibits the most activity, as many other stocks generate predictions of 0 or -1, resulting in no buy actions.

Table 30. TOBA.JK Trading Simulation for 2 Months

Number	Prediction	Action	Lot	Avg. Cost	Open Price	Investment	Close Price	Return	Gain/Loss
1-7	0	Stay			Rp358.00		Rp358.00		
8	1	Buy	50	Rp358.00	Rp358.00	Rp1,790,000.00	Rp358.00	0.00%	
9-11	0	Stay	50	Rp358.00	Rp354.00	Rp1,770,000.00	Rp354.00	1.12%	
12	1	Buy	50	Rp358.00	Rp354.00	Rp3,520,000.00	Rp352.00	-1.68%	
			50	Rp354.00				-0.56%	
13	1	Buy	100	Rp356.00	Rp352.00	Rp5,280,000.00	Rp352.00	-1.12%	
			50	Rp352.00				0.00%	
14-17	0	Stay	150	Rp354.00	Rp370.00	Rp5,550,000.00	Rp370.00	4.52%	
18	-1	Sell	150	Rp354.00	Rp370.00	Rp5,550,000.00	Rp370.00	4.52%	Rp240,000.00
19-26	0	Stay			Rp350.00		Rp350.00		
27	1	Buy	50	Rp350.00	Rp350.00	Rp1,750,000.00	Rp350.00	0.00%	
28-30	0	Stay	50	Rp350.00	Rp348.00	Rp1,730,000.00	Rp346.00	-1.14%	
31	1	Buy	50	Rp350.00	Rp346.00	Rp3,460,000.00	Rp346.00	-1.14%	
			50	Rp346.00				0.00%	
32	1	Buy	100	Rp348.00	Rp346.00	Rp5,250,000.00	Rp350.00	0.57%	
			50	Rp346.00				1.16%	
33	1	Buy	150	Rp347.00	Rp350.00	Rp7,000,000.00	Rp350.00	0.86%	
			50	Rp350.00				0.00%	
34	1	Buy	200	Rp348.50	Rp350.00	Rp8,750,000.00	Rp350.00	0.43%	
			50	Rp350.00				0.00%	
35	1	Buy	250	Rp349.25	Rp350.00	Rp10,500,000.00	Rp350.00	0.21%	
			50	Rp350.00				0.00%	
36	1	Buy	300	Rp349.63	Rp350.00	Rp12,250,000.00	Rp350.00	0.11%	
			50	Rp350.00				0.00%	
37-40	0	Stay	350	Rp349.81	Rp370.00	Rp12,950,000.00	Rp370.00	5.77%	
42	-1	Sell	350	Rp349.81	Rp370.00	Rp12,950,000.00	Rp370.00	5.77%	Rp706,562.50

The generated model has achieved a sufficiently high level of accuracy, as evident in the stock simulation, which also yielded positive results. The outcome is a profit of Rp 946,562.50 with a final return of 5.77%. This result indicates that the investment decisions generated by the model were successful, at least in this simulation scenario.

V. CONCLUTIONS AND RECOMMENDATIONS

A. CONCLUTIONS

In the research, the optimized random forest method has been applied to analyze stock trends in the energy sector. Two models were constructed based on two types of data: stock-based data and annual data. Each model development involves four stages: random forest model, optimized random forest model, random forest model with feature importance, and optimized random forest model with feature importance. The results were obtained through a comparison of the two model developments.

The results based on stock-based data indicate that the optimized random forest model with feature importance and the random forest model with feature importance excels in both datasets. On the other hand, the results based on year-based data show that the optimized random forest method outperforms the random forest method with three datasets. When comparing all data types and their models, it is found that the optimized random forest method is more effective in predicting stock trends compared to the random forest method. Through simulations, it was also discovered that with this model, the selected stocks resulted in a profit of Rp 946,562.50 or with a final return of 5.77%.

B. RECOMMENDATIONS

Through the conducted research, it can be observed that the optimized random forest model is capable of predicting stock price trends in Indonesia. However, it has been proven from the results obtained in stock market simulations that the model still incurs losses. Here are some suggestions for future researchers:

1. Utilize stock data from other sectors.
2. Partition the data into a training set and a test set. For imbalanced data, consider using alternative methods to produce more diverse results.
3. Explore other methods such as Support Vector Machine and Neural Network.
4. Experiment with different return close price limits and time periods.
5. Incorporate other technical indicators with high correlations to the data to enhance prediction accuracy.

REFERENCES

[1] T. Darmadji and H. M. Fakhruddin, Pasar Modal di Indonesia: Pendekatan Tanya Jawab, Salemba Empat, 2006.

[2] S. Rahardjo, Kiat Membangun Aset Kekayaan, Elex Media Komputindo, 2006.

[3] H. A. Al Hakim, "Prediksi tren pergerakan harga saham menggunakan algoritma temporal convolutional network (tcn)," 2021.

[4] L. Yin, B. Li, P. Li and R. Zhang, "Research on Stock Trend Prediction Method Based on Optimized Random Forest," *CAAI Transactions on Intelligence Technology,* pp. 274 - 284, 2021.

[5] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo and M. Chica-Rivas, "Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machine," *Ore Geology Reviews,* no. 71, pp. 804-818, 2015.

[6] A. Hermansyah, "Analisis teknikal pergerakan harga saham untuk mengambil keputusan investasi pada saham sub sektor telekomunikasi yang terdaftar di bursa efek indonesia," 2020.

[7] Z. B. Zabinsky and Others, "Random Search Algorithms," *Department of Industrial and Systems Engineering,* 2009.

[8] S. Basak, S. Kar, S. Saha, L. Khaidem and S. R. Dey, "Predicting the direction of stock market prices using tree-based classifiers," *The North American Journal of Economics and Finance,* no. 47(C), pp. 552-567, 2019.

[9] M. Ballings, D. Van den Poel, N. Hespeels and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.,* no. 42(20), pp. 7046-7056, November 2015.

[10] D. Berrar, "Cross-Validation," January 2018.

[11] E. Ostertagova and O. Ostertag, "Forescasting using simple exponential smoothing method," *Acta Electrotechnica et Informatica,* no. 12, pp. 62-66, December 2012.

[12] L. Breiman, "Random forests," *Machine learning,* no. 45(1), pp. 5-32, 2001.

[13] C. Nguyen, Y. Wang and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," 2013.

[14] L. Rokach and O. Maimon, *Decision Trees,* vol. 6, pp. 165-192, January 2005.

[15] javaTpoint, "Decision Tree Classification Algorithm," [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm. [Accessed 26 October 2022].

[16] S. Visa, B. Ramsay, A. L. Ralescu and E. Van Der Knaap, "Confusion matrix-based feature selection," *MAICS,* no. 710(1), pp. 120-127, 2011.

[17] "Understanding confusion matrix," [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62. [Accessed 11 October 2022].

[18] L. Fu, P. Liang, X. Li and C. Yang, "A machine learning based ensemble method for automatic multiclass classification of decisions," *Evaluation and Assesment in Software Engineering,* pp. 40-49, 2021.

[19] A. Evered, "Attentional and perceptual processes in a medical image interpretation task," March 2016.

[20] P. Refaeilzadeh, L. Tang and H. Liu, "Cross-validation," *Encyclopedia of database systems,* no. 5, pp. 532-538, 2009.