MethodsX 13 (2024) 103031

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Ocean wave prediction using Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) in Tuban Regency for fisherman safety ☆,☆☆

Riswanda Ayu Dhiya'ulhaq, Anisya Safira, Indah Fahmiyah, Mohammad Ghani*

Data Science Technology, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Surabaya 60115, Indonesia

ARTICLE INFO

Method name:
Long Short-Term Memory and Extreme Gradient Boosting

Keywords:
Long Short-Term Memory (LSTM)
Extreme Gradient Boosting (XGBoost)
Wave prediction

ABSTRACT

The fishing industry has a large role in the Indonesian economy, with potential profits in 2020 of around US$ 1.338 billion. Tuban Regency is one of the regions in East Java that contributes to the fisheries sector. Fisheries relate to the work of fishermen. Accidents in shipping are still a major concern. One of the natural factors that influence shipping accidents is the height of the waves. Fisherman safety regulations have been established by the Ministry of Maritime Affairs and Fisheries and the Meteorology, Climatology and Geophysics Agency. Apart from regulations, the results of wave height predictions using the Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) methods can help fishermen determine shipping departures, thereby reducing the risk of accidents. In this study, the Grid Search hyperparameter tuning process was used for both methods which were carried out on four location coordinates. Based on the analysis results, LSTM is superior in predicting wave height for the next 30 days because it can predict wave height at all three locations, with results at the first location (RMSE 0.045; MAE 0.029; MAPE 8.671 %), second location (RMSE 0.051; MAE 0.035; MAPE 10.64 %), and third location (RMSE 0.044; MAE 0.027; MAPE 7.773 %), while XGBoost only has the best value at fourth location (RMSE 0.040; MAE 0.025; MAPE 7.286 %).

- Hyperparameter tuning with gridsearch is used in LSTM and XGBoost to obtain optimal accuracy
- LSTM outperforms in three locations, while XGBoost outperforms in the fourth location.
- Advanced prediction techniques such as LSTM and XGBoost improve fishermen's safety by providing accurate wave height estimates, thereby reducing the possibility of shipping accidents.

Specification table

Subject area:	Engineering
More specific subject area:	Ocean Engineering
Name of your method:	Long Short-Term Memory and Extreme Gradient Boosting
Name and reference of origin method:	None
Resource availability:	https://github.com/riswandayu/predictoceanwave

☆ **Related research article:** None.
☆☆ **For a published article:** None.

* Corresponding author.
E-mail address: mohammad.ghani@ftmm.unair.ac.id (M. Ghani).

https://doi.org/10.1016/j.mex.2024.103031
Received 26 July 2024; Accepted 1 November 2024
Available online 2 November 2024
2215-0161/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/)

Background

Fishing is inherently hazardous, characterized by its 3D characteristics dangerous, dirty, and difficult [1]. Among the natural elements that affect maritime activities, wave height is the most prominent. The Meteorology, Climatology, and Geophysics Agency (BMKG) provide important safety advice, recommending that fishermen navigate when the wave height does not exceed 1.25 m [2]. Wave height estimation plays a crucial role in determining the safe departure of fishing vessels. Various methodologies, including stochastic techniques and deep learning, are commonly used. Stochastic methods such as Autoregressive Integrated Moving Average (ARIMA) predict wave patterns, especially in Jakarta Bay. ARIMA experiments involving eight combinations yielded an optimized Root Mean Square Error (RMSE) value for the ARIMA (2,2,2) model of 0.0106, indicating its ability to accurately predict waves up to 24 h in advance [3]. ARIMA performs well for short-term forecasting it tends to have more significant errors when it comes to long-term predictions. Furthermore, since ARIMA relies on historical data it may not fully account for external factors naturally occurring in the real world [4]. Another interesting method is Random Forest (RF), which is widely used in operational wave forecasting in the Atlantic Ocean. Campos et al. [5] highlighted the decline in RF performance as the forecast time increases, RF is only able to forecast in very short timescales, thus the prediction results become uncertain with the long term forecast, indicating its limitations in long term prediction.

Londhe and Panchang [6] applied artificial neural networks (ANNs) using a feed-forward backpropagation algorithm to forecast significant wave heights at four different lead times (6,12, 18, and 24 h). This study was conducted across six buoy locations in the Gulf of Mexico, Alaska, and Maine. They tested six different network architectures for each buoy site, generating predictions from January 8 to December 31, 2004, for all buoys except one, which had predictions that concluded on September 16 of that year. The results showed an accuracy of 86 % for the 6-hour lead time and between 67 % and 83 % for the 12-hour lead time. As anticipated, the accuracy decreased for longer lead times, dropping to between 55 % and 71 % for 18 h and falling below 63 % for the 24-hour forecasts. Despite these promising results, a noted limitation of using ANNs for ocean wave predictions was the significant under-predictions of the highest wave peaks, which could be attributed to the prevalence of lower wave height records in the datasets used for training the network [7]. The traditional architecture of artificial neural networks (ANNs) has undergone improvements, particularly with the introduction of recurrent neural networks (RNNs), which incorporate at least one feedback loop to create cyclic connections between layers. This architecture aims to enhance prediction accuracy by allowing the network to retain memory from one time step to the next. However, the effectiveness of RNNs in capturing contextual information is still constrained by the vanishing gradient problem [7]. As information circulates within the recurrent network over time, the influence of an input on the hidden layer, and consequently on the output, can either decay or escalate exponentially [7]. Similar to the ANN models, RNNs utilize the backpropagation algorithm for learning; however, this method often proves ineffective due to the gradient values either exploding (increasing significantly) or vanishing (decreasing to near zero). As the number of iterations (epochs) increases, these gradient values may cause the model to lose its ability to learn [8].

The method to overcome the ANN problem is Long Short-Term Memory (LSTM), a modification of the Recurrent Neural Network (RNN) capable of processing, predicting, and classifying information based on time series data. LSTM networks have a significant advantage over RNNs, particularly due to the presence of the forget gate. The forget gate is a component that functions to determine which information from the historical data should be retained or discarded from the LSTM memory cell [9]. LSTM performs faster than RNN because it can add its own loop method. LSTM is an effective method for handling sequential data, making it a valuable resource for scientific research across various disciplines, such as oceanography [9]. Meng at al [10] explained that the application of LSTM in predicting significant wave heights has superior accuracy compared to other methods. Research on predicting significant wave heights and peak wave periods used for wave energy converters (WECs) with a dataset spanning two years showed significant results in predicting wave periods for the next 3 h with an RMSE value of 0.1260. Meanwhile, with a dataset spanning four years, LSTM produced better results for the next 6 h, with an RMSE of 0.0268. The research conducted by [10] used wave data from Wave Analysis for Fatigue and Oceanography (WAFO) and found that LSTM achieved good accuracy when using 100 inputs, as evidenced by the minimal MAE and RMSE values of 0.015 and 0.019, respectively. Another study in the Bali Strait, Indonesia [11], showed that LSTM performs well in predicting the U component (east-west current velocity) with a MAPE of 18.64 % and the V component (north-south current velocity) with a MAPE of 5.29 %. Another study in Indonesian waters was conducted by Abdullah et al. [8] where data were collected from four stations: the first station in Baron, an open sea area; the second in Belawan, a strait; the third in Karawang, with nearshore wave conditions; and the fourth in Masalembo, a deep-sea area. The research employed two methods: single-step and multi-step. The results showed that LSTM can generate accurate hourly predictions. Meanwhile, for the multi-step scheme, LSTM's performance in predicting wave heights over 12, 24, and 48-hour periods demonstrated that the model could effectively capture trend data [8].

Another method widely used for predicting wave heights is XGBoost. XGBoost, short for Extreme Gradient Boosting, is a machine learning algorithm introduced by Chen and Guestrin [12]. XGBoost is an implementation of the gradient boosting algorithm. Research conducted by Hu et al. [13] on wave height prediction in Lake Erie produced a MAPE of 23.4 %–30.8 % for LSTM, while XGBoost performed better with a MAPE of 16.6 %–22.9 %. In wave height prediction research conducted in Pangandaran, Indonesia [14], XGBoost was found to be more optimal than AdaBoost, with the best RMSE value of 0.064 for 14-day predictions. XGBoost was also used to predict wave heights on flat shores, resulting in an MAE of 0.026 and an RMSE of 0.039 [15]. Proper hyperparameter tuning can improve RMSE by 11 % [16]. The LSTM method is widely used for long-term predictions (over 100 s) of irregular regular wave elevation. The LSTM model is used to predict future wave profiles based on historical wave data. The capability of LSTM to understand long-term dependencies in oceanographic data has allowed it to produce precise wave height predictions, which are vital for ensuring marine safety and aiding in coastal planning. By utilizing input, forget, and output gates to manage the flow of information, the LSTM

Fig. 1. Coordinate map.

model can effectively identify intricate patterns and dependencies [9]. Meanwhile, XGBoost has the ability to handle missing data, as well as large and complex data sets. Additionally, the performance of LSTM for wave prediction takes 0.24 s on 1 CPU, while XGBoost only requires 0.03 s on 1 CPU, making it faster compared to physics-based wave prediction methods [13]. Based on this explanation, this research will use LSTM and XGBoost to predict sea wave heights in Tuban Regency, with the hope that these predictions will help reduce the risk of accidents, minimize losses, and save fishermen's lives. This research aligns with the achievement of point 13 of the Sustainable Development Goals (SDGs), which emphasizes addressing climate change. Therefore, these efforts not only protect the safety of fishermen but also align with the global vision of tackling climate change challenges.

Method details

Case study and data source

Maritime accidents In Tuban Regency, have significantly impacted local fishing communities between 2020 and 2023. In 2020, seven fishing vessels sank at the dock in Palang District, and another vessel in Tambakboyo District suffered losses of IDR 200 million due to high waves. The trend continued with three vessels sinking in 2021, two vessels in 2022 resulting in two missing fishermen, and in May 2023, a vessel capsized due to high waves, causing three fatalities. These incidents highlight the vulnerability of small-scale fishing operations to severe weather conditions, emphasizing the need for improved safety measures and early warning systems in the region.

The data source used in this research is secondary data obtained from Perak Meteorological Station Surabaya II. Wave height data is used for one year, starting from November 2022 to October 2023, with the water area taken from the coordinates −6.21˚S;111.80˚E to −6.31˚S;112.27˚E as in Fig. 1. The determination of the location point is based on the activities of fishermen and also the fishery products that are more abundant when the waters are 20 to 30 miles from the coast [17]. There are four observed locations, the first location with coordinates −6.21˚S;111.80˚E, the second location with coordinates −6.22˚S;112.04˚E, the blue line shows the third location with coordinates −6.26˚S;112.20˚E, and the green line shows the fourth location with coordinates −6.31˚S;112.27˚E. This data includes time-related information containing data per day with time every 6 h starting at 00.00, 06.00, 12.00, and 18.00, longitude values, latitude values, and High Significant Wave (HSW) parameters with a total of 5348 lines. Regarding data availability, although the dataset is comprehensive, challenges may arise related to spatial and temporal resolution, as well as potential inconsistencies in data collection. Any missing or incomplete records may introduce uncertainty, potentially affecting the reliability of the models developed in this study. In addition, the selection of sites based on fishing activity may introduce bias, limiting the generalizability of the results to other areas with different wave conditions

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) was introduced by Hochreiter and Schmidhuber in 1997 [18]. LSTM is one part of deep learning which is a subtype of Recurrent Neural Networks (RNN). LSTM is a modification of RNN that has the advantage of solving the missing gradient problem [19]. LSTM can learn and handle very long time dependencies, with a minimum time lag of >1000 time steps [20]. The LSTM has four interacting layers [21] where the LSTM architecture is shown in [22]. The gates in an LSTM cell can be explained as follows:

$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right) \tag{1}$$

$$f_t = \sigma\left(x_t U^f + h_{t-1} - W^f\right) \tag{2}$$

$$o_t = \sigma\left(x_t U^o + h_{t-1} - W^o\right) \tag{3}$$

Table 1
LSTM hyperparameter tuning.

Parameters	
Hidden Layer	4
Optimizer	Adam
Neuron	30;50
Activation Function	Tanh
Learning Rate	0,001; 0,01; 0,1
Batch Size	32;64
Epoch	50; 100

Table 2
XGBoost hyperparameter tuning.

Parameters	
colsample_bytree	0,5; 0,7; 1
learning_rate	0,05; 0,15; 0,3
max_*depth*	3; 5; 6
n_estimators	50; 100; 150
reg_lambda	0; 0,1; 0,5; 1
Subsample	0,6; 1

$$\widetilde{C_t} = \tanh\left(x_t U^g + h_{t-1} W^g\right) \tag{4}$$

$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \widetilde{C_t}\right) \tag{5}$$

$$h_t = \tanh(C_t) * o_t \tag{6}$$

where i_t is the input gate, f_t is the forget gate, o_t is the output gate, W is the connection between the current and previous nodes, U contains the weights of the inputs to the *hidden layer,* $\widetilde{C_t}$ is candidate *hidden layer,* and C is the memory unit.

LSTM hyperparameter tuning

The LSTM hyperparameter tuning method uses Grid Search, which is defined as a method of testing a given hyperparameter combination on a grid configuration [23]. The grid search method will search for all possibilities by preparing a grid, which is then evaluated to get the best value among all grids [24]. Table 1 presents the parameters used in LSTM. In this research, the Tanh activation function and ADAM optimizer are used. The calculation of the Tanh gradient is simple and converges faster than sigmoid [25]. Adam is an adaptive learning rate technique that lowers the individual learning rate for various parameters [26].

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm introduced by Chen and Guestrin in 2016 [12]. XGBoost is one of the boosting methods. Boosting is an ensemble learning technique that combines a set of simpler and weaker sets to improve model accuracy [27]. XGBoost calibrates the previous predictor using the residual. Loss function optimization (LOF) is involved in this process [27]. However, regularization is applied to the objective function during calibration to reduce overfitting [28]. XGBoost minimizes regularized losses, and use of L2 regularization aims to obtain a more general model. L2 regularization helps prevent overfitting, adding L2 regularization in the loss function results in more controllable model complexity [29]. In general, the L2 regularization equation is as follows:

$$Cost\ Function = loss + \frac{\lambda}{2m} * \sum \|w\|^2 \tag{7}$$

where λ is the L2 regularization parameter whose value can be optimized for better results, m the number of trees and w is the weight vector or model parameters.

XGBoost hyperparameter tuning

XGBoost is a highly promising machine learning algorithm with numerous hyperparameters [30]. A model's hyperparameters should be selected optimally to prevent overfitting and underfitting. These hyper-parameters are learned and adjusted based on the data to achieve the most suitable fit [31]. The XGBoost hyperparameter tuning method uses Grid Search. Boosting reduces bias and increases variance by increasing the complexity of weak models. By using hyperparameters, overfitting can be prevented. Table 2 describes the hyperparameters used in XGBoost.

Effectively, these hyperparameters affect the tree building and Gain calculation (λ and γ) or the data and feature selection process for each iteration such as subsample and colsample bytree, by adjusting them it can maintain a balance between model complexity and predictive ability, and improve performance by reducing overfitting [32].

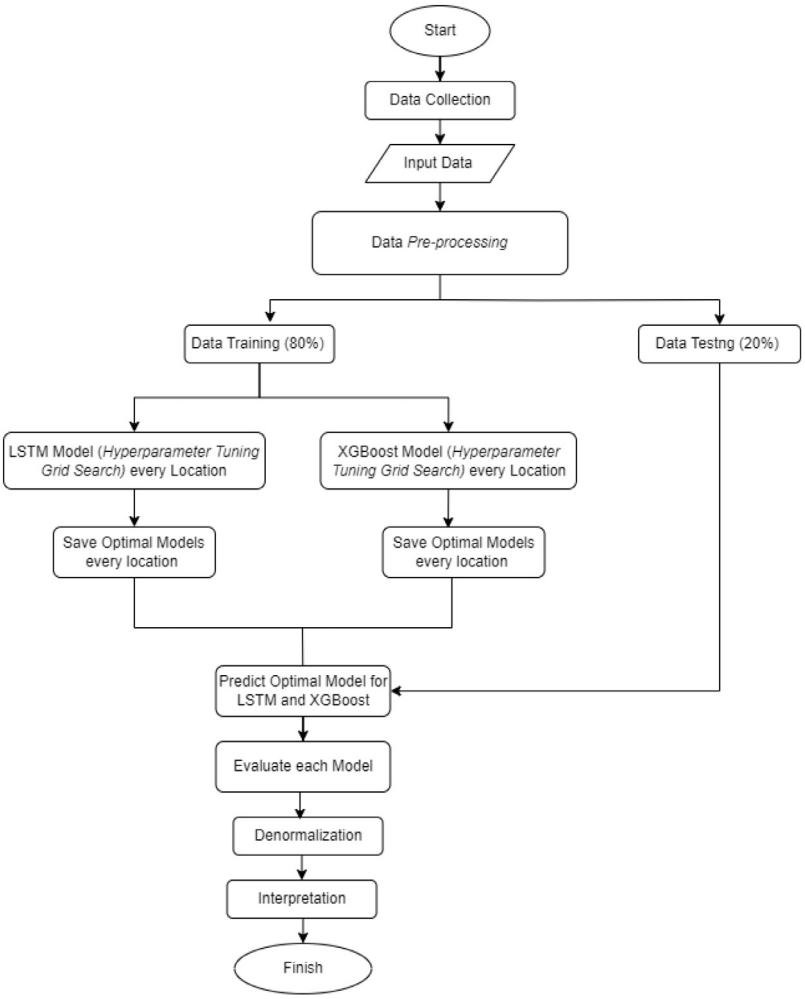

Fig. 2. Flowchart.

Prediction flowchart with LSTM and XGboost

The research work is divided into two parts, the prediction part with LSTM and XGBoost described in Fig. 2.

1. Input Data.
2. Data preprocessing, filtering data based on location and normalizing data. The method used in the normalization process is min-max scaler by converting the actual value into a value with an interval range [33]. The following is the normalization formula.

$$X' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{8}$$

where X' is the normalized value, x is the actual data value to be normalized, x_{min} is the minimum value of the actual data, and x_{max} is the maximum value of the actual data. It is important to note that this normalization process only adjusted the data scale, as no missing values were found in the dataset. Additionally, no outlier removal was conducted since the outliers contain critical information regarding extreme wave events.

3. Data division with a percentage of 80 % for training data and 20 % for testing data.
4. LSTM modeling for each location using hyperparameter tuning with grid search and using early stopping to prevent overfitting or underfitting [34].
5. Modeling XGBoost for each location using hyperparameter tuning with grid search. The best parameter tuning results at each location will be applied in building a prediction model for each location. In grid search, negative square error scoring is used.
6. Model evaluation to see the performance of the model using RMSE, MAE, and MAPE. The selection of RMSE, MAE, and MAPE as performance metrics was driven by their ability to provide clear and interpretable insights into model accuracy, particularly in the context of time series forecasting. RMSE and MAE offer straightforward measures of error magnitude [35] with RMSE being particularly sensitive to larger errors [36], which is valuable in identifying significant deviations. MAPE, on the other hand,

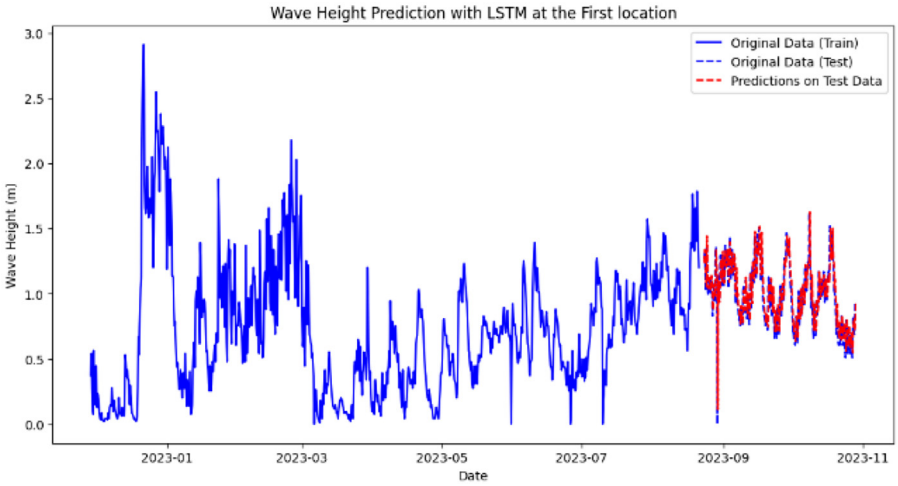

Fig. 3. Comparison of actual and predicted values location 1.

expresses errors as a percentage, allowing for easy comparison across different scales [11]. Here is the evaluation formula [37].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{11}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, dan n is the number of predictions.

7. Denormalization, to return the predicted value to its original value. Here is the formula for denormalization [38].

$$d = y'(\max(y) - \min(y)) + \min(y) \tag{12}$$

d is the denormalized value, y' is the normalized value, max is the minimum value of the actual data, and min is the maximum value of the actual data.

8. Model Interpretation, Interpretation is related to the visualization of data presentation.

Method validation

Long Short-Term Memory (LSTM)

1st Location (−6.21°S;111.80°E)

Based on the tuning results using grid search for the first location, the negative mean square error value of −0.00352 or RMSE value of 0.05939 is the best parameter using batch size 32; epoch 50; activation tanh; learning rate 0.01; number of neurons 50; optimizer Adam. The best parameters are then used to make wave height predictions, obtained MAPE evaluation results of 8.671 %; MAE 0.029; and RMSE 0.045 for the first location. The prediction of wave height at the first location (Fig. 3) shows that between the actual value and the predicted value has a relatively small difference, the difference does not look significant.

Based on the evaluated model, a prediction of the wave height for the next 30 days was made using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 4. The green line shows the pattern of the predicted wave height for the next 30 days, which tends to decrease when compared to the previous period.

2nd Location (−6.22°S;112.04°E)

The tuning results using grid search for the second location obtained a negative mean square error value of −0.00290 or RMSE value of 0.05385 with parameters batch size 64; epoch 100; activation tanh; learning rate 0.001; number of neurons 50; optimizer Adam. The best parameters are then used to make wave height predictions, obtained MAPE evaluation results of 10.79 %; MAE 0.035; and RMSE 0.051 for the second location. The prediction of wave height at the second location (Fig. 5) shows that the prediction plot (red line) has shifted slightly to the right but there is no significant difference between the actual value and the predicted value, it can be seen from the pattern between the two lines that have almost the same pattern and height.

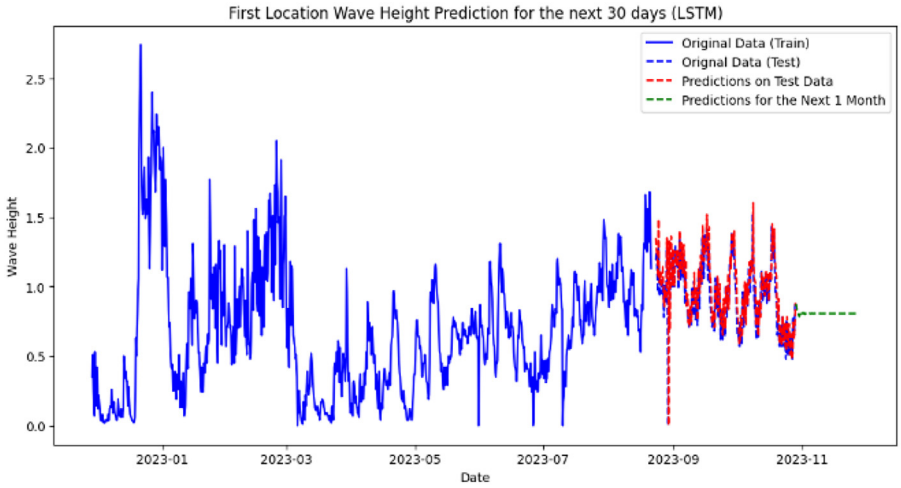

Fig. 4. Next 30-day wave height prediction.

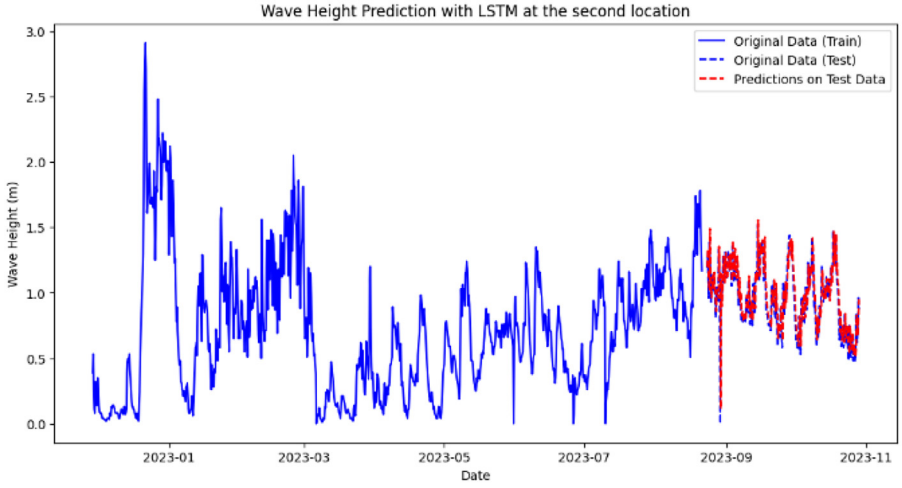

Fig. 5. Comparison of actual and predicted values location 2.

Based on the evaluated model, the wave height prediction for the next 30 days was then carried out using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 6. The green line shows the pattern of the wave height prediction for the next 30 days, which tends to decrease when compared to the previous period.

3rd Location (−6.26°S;112.20°E)

The tuning results using grid search for the third location obtained a negative mean square error value of −0.00214 or RMSE value of 0.04633 with the best parameters using batch size 32; epoch 50; activation tanh; learning rate 0.001; number of neurons 50; optimizer Adam. The best parameters are then used to make wave height predictions, obtained MAPE evaluation results of 7.773 %; MAE 0.027; and RMSE 0.044 for the third location. Wave height prediction at the third location (Fig. 7) shows that the prediction plot (red line) at the third location tends to have a lower value when compared to the actual data.

Based on the evaluated model, a prediction of the wave height for the next 30 days was then made using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 8. The green line shows the pattern of the predicted wave height for the next 30 days, which tends to decrease when compared to the previous period.

4th Location (−6.31°S;112.27°E)

The tuning results using grid search for the fourth location obtained a negative mean square error value of −0.00203 or RMSE value of 0.04506 with the best parameters using batch size 32; epoch 100; activation tanh; learning rate 0.001; number of neurons 30; optimizer Adam. The best parameters were then used to make wave height predictions, obtained MAPE evaluation results of 7.386 %; MAE 0.025; and RMSE 0.042 for the fourth location. Wave height prediction at the fourth location (Fig. 9) shows the prediction plot

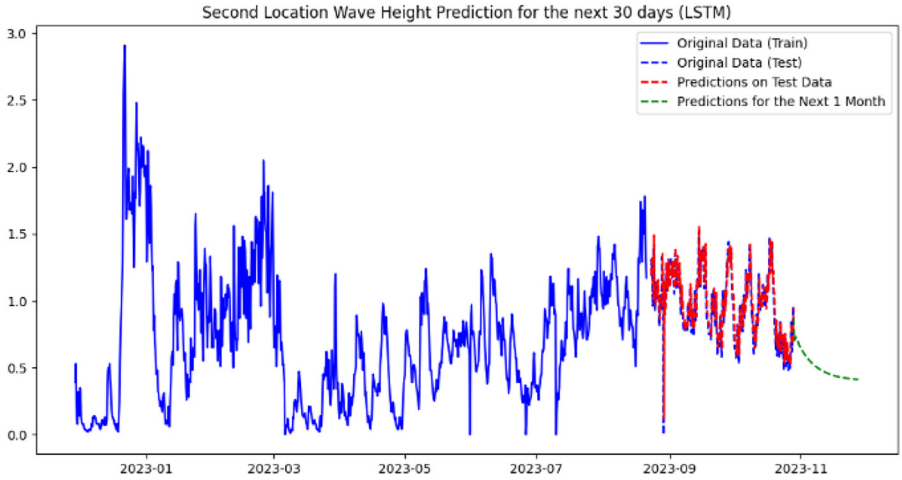

Fig. 6. Next 30-day wave height prediction.

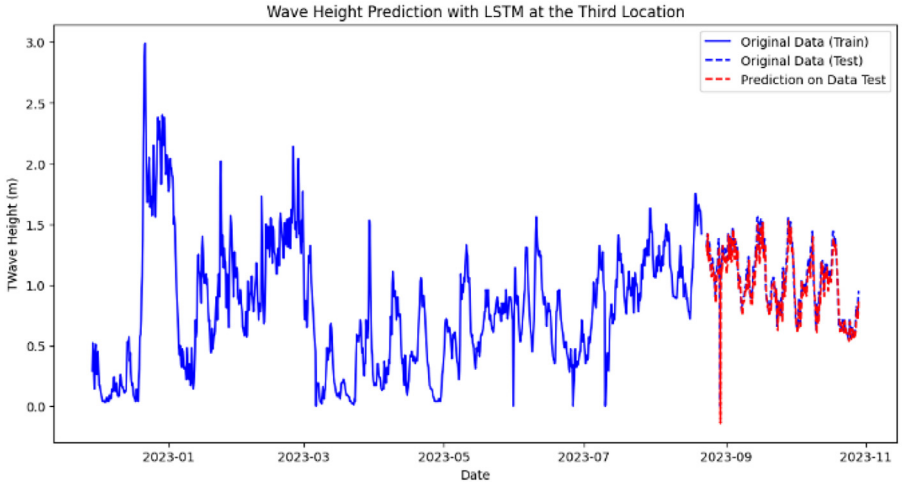

Fig. 7. Comparison of actual and predicted values location 3.

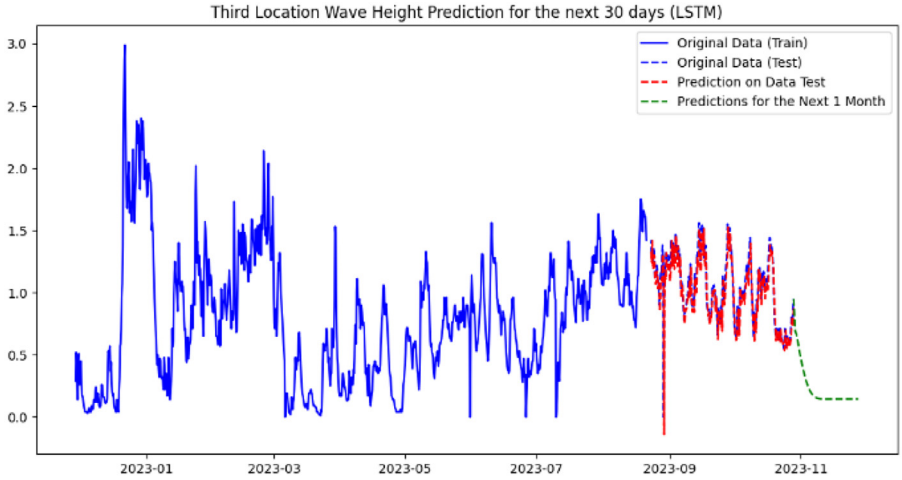

Fig. 8. Next 30-day wave height prediction.

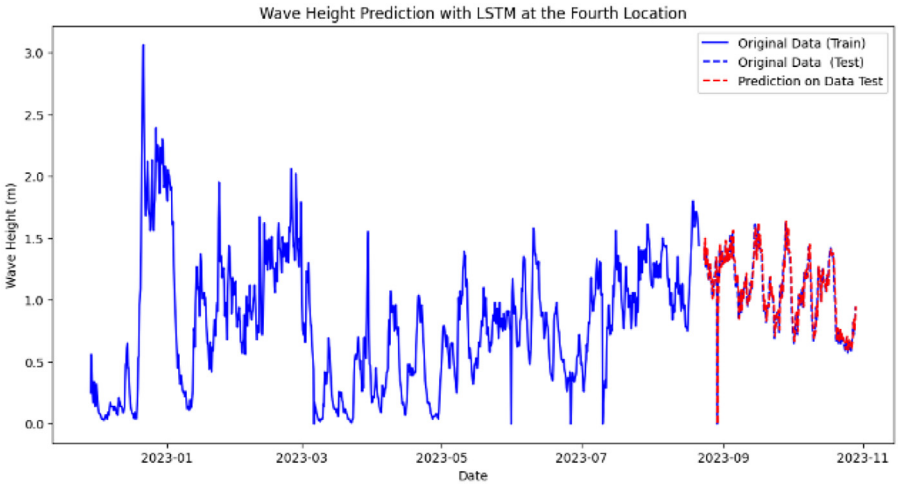

Fig. 9. Comparison of actual and predicted values location 4.

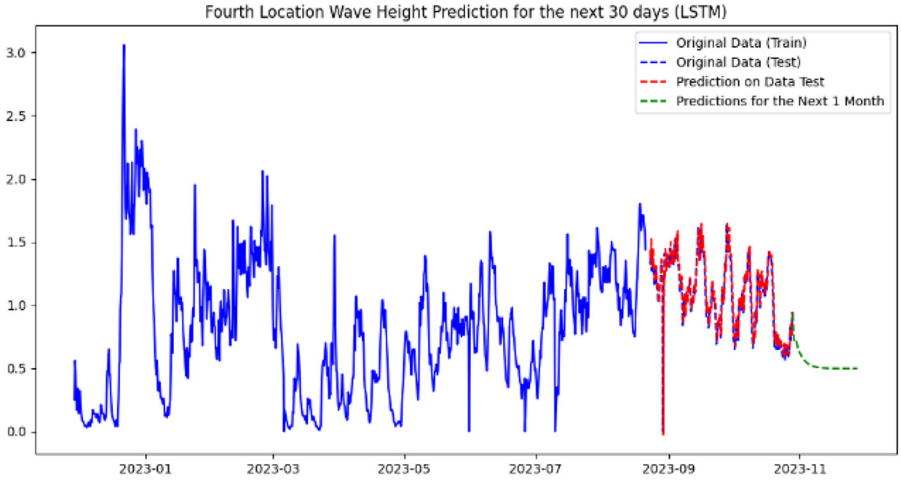

Fig. 10. Next 30-day wave height prediction.

(red line) tends to follow the pattern and height of the actual data, when compared to the visualization results of the other three locations, the prediction plot on the fourth location test data is more fit to the actual data.

Based on the evaluated model, a prediction of the wave height for the next 30 days was made using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 10. The green line shows the pattern of the predicted wave height for the next 30 days, which tends to decrease when compared to the previous period.

Extreme Gradient Boosting (XGBoost)

1st Location (−6.21°S;111.80°E)

By using grid search, the best parameter combination is colsample bytree 1; learning rate 0.05; max depth 3; n estimator 150; reg lambda 0.1; subsample 0.6 with a negative mean square error value of −0.00406 or RMSE value of 0.06376. The best parameters are then used to make wave height predictions, obtained MAPE evaluation results of 9022 %; MAE 0.031; and RMSE 0.046 for the first location. The prediction results show that the test data (blue line) (Fig. 11) has a higher value when compared to the prediction plot, however the prediction plot pattern is quite following the pattern of the actual data.

Based on the evaluated model, a prediction of the wave height for the next 30 days was made using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 12. The green line shows the pattern of the predicted wave height for the next 30 days, which tends to decrease when compared to the previous period.

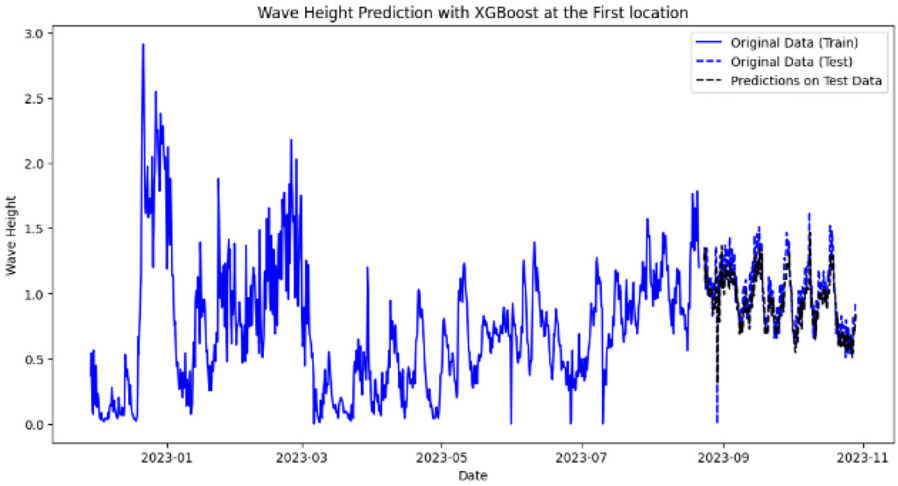

Fig. 11. Comparison of actual and predicted values location 1.

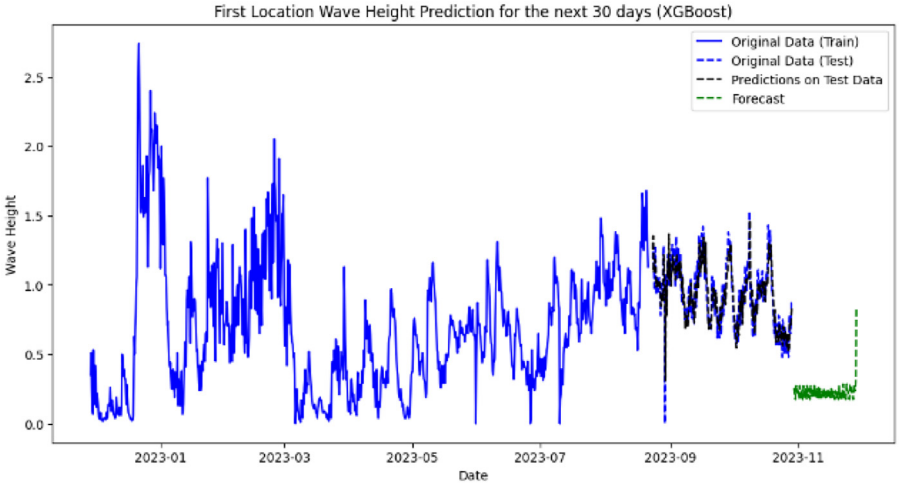

Fig. 12. Next 30-day wave height prediction.

2nd Location (−6.22°S;112.04°E)

The best parameter combination obtained with grid search is colsample bytree 1; learning rate 0.15; max depth 3; n estimator 100; reg lambda 0.5; subsample 0.6 with a negative mean square error value of −0.00374 or RMSE value of 0.06122. The best parameters are then used to make wave height predictions, obtained MAPE evaluation results of 11.05 %; MAE 0.036; and RMSE 0.050 for the second location. A visualization of the prediction results is presented in Fig. 13. The prediction plot (black line) has higher values at the end of August and mid-September when compared to the actual values. Overall, it can be observed that the plot of the predicted data does not follow the pattern of the actual data.

Based on the evaluated model, the wave height prediction for the next 30 days was then carried out using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 14. The green line shows the pattern of the wave height prediction for the next 30 days, which tends to decrease when compared to the previous period.

3rd Location (−6.26°S;112.20°E)

The best parameter combination for the third location is colsample bytree 1; learning rate 0.15; max depth 3; n estimator 50; reg lambda 0; subsample 0.6 with a negative mean square error value of −0.00267 or an RMSE value of 0.05171. The best parameters are then used to make wave height predictions, obtained MAPE evaluation results of 8160 %; MAE 0.028; and RMSE 0.043 for the third location. Based on Fig. 15, the prediction plot (black line) has a higher value in early September when compared to the actual value. Then in October the prediction plot also shows lower values than the actual data. However, overall the plots of the actual and predicted data have almost the same pattern and height.

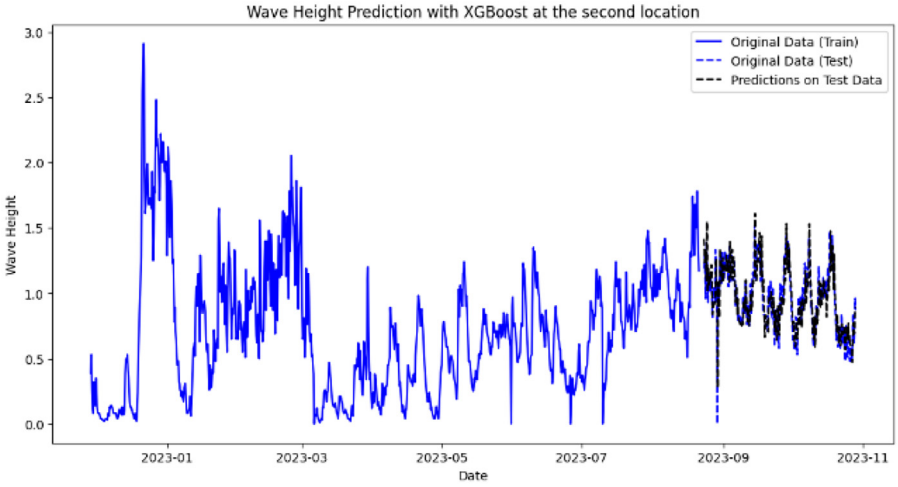

Fig. 13. Comparison of actual and predicted values location 2.

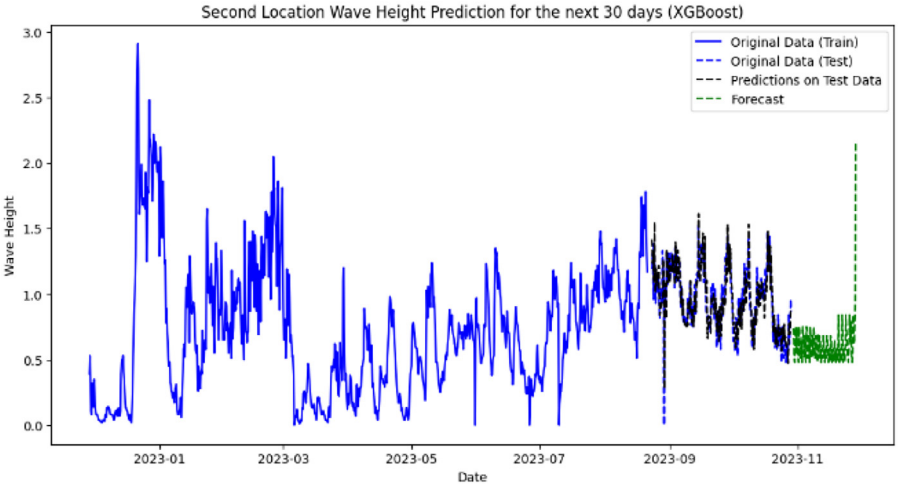

Fig. 14. Next 30-day wave height prediction.

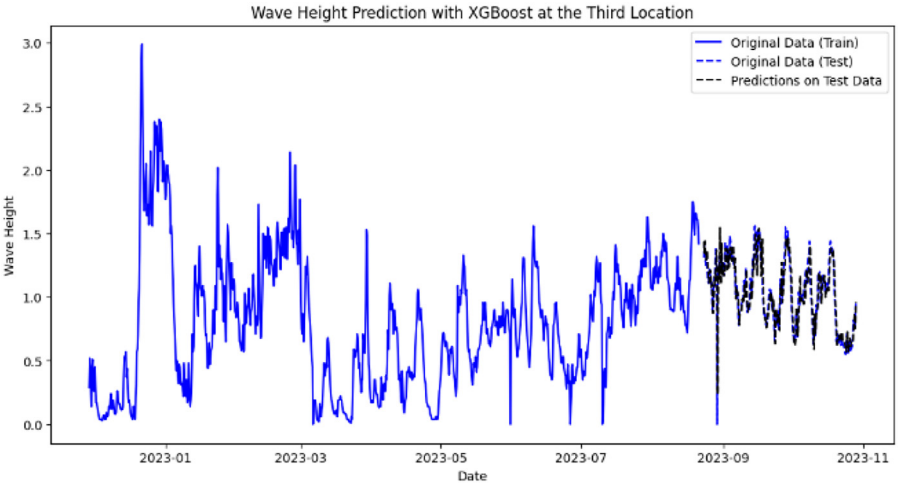

Fig. 15. Comparison of actual and predicted values location 3.

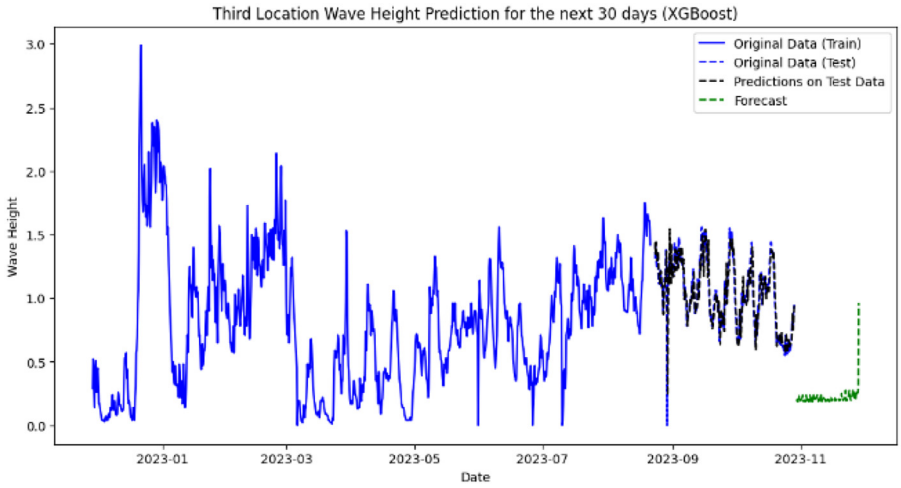

Fig. 16. Next 30-day wave height prediction.

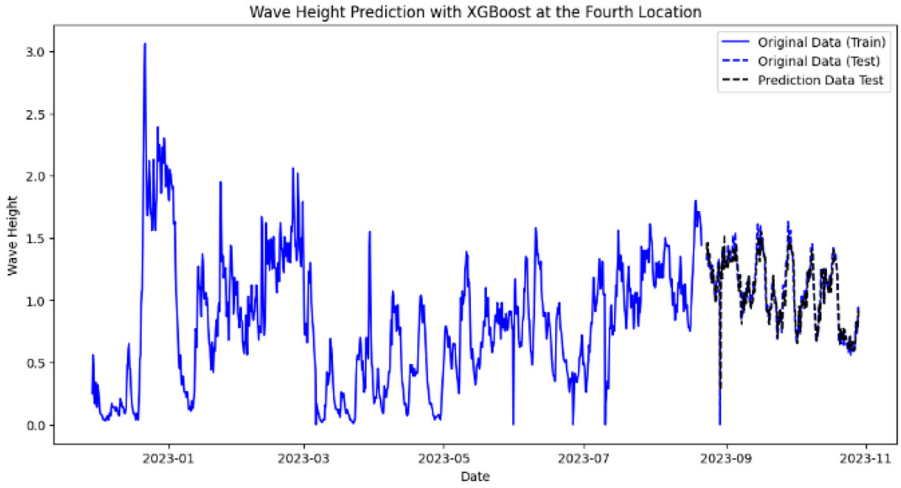

Fig. 17. Comparison of actual and predicted values location 4.

Based on the evaluated model, a prediction of the wave height for the next 30 days was then made using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 16. The green line shows the pattern of the predicted wave height for the next 30 days, which tends to decrease when compared to the previous period.

4^{th} Location (−6.31°S;112.27°E)

The best parameter combination for the fourth location is colsample bytree 1; learning rate 0.05; max depth 3; n estimator 100; reg lambda 0; subsample 1 with a negative mean square error value of −0.00275 or an RMSE value of 0.05249. The best parameters were then used to make wave height predictions, obtained MAPE evaluation results of 7.286 %; MAE 0.025; and RMSE 0.040 for the fourth location. Based on Fig. 17 the prediction plot (black line) has a higher value in early September when compared to the actual value. In October the actual data plot tends to be higher when compared to the prediction data plot.

Based on the evaluated model, a prediction of the wave height for the next 30 days was made using the historical data in the test data. The prediction is done to predict the wave height at the first location every 6 h, which is presented in Fig. 18. The green line shows the pattern of the predicted wave height for the next 30 days, which tends to decrease when compared to the previous period.

Comparison of LSTM and XGBoost prediction results

The comparison results of the two models can be used to determine the effectiveness of the model in predicting wave heights in the Tuban Regency with four location coordinates. The comparison results of the two models are presented in Table 3

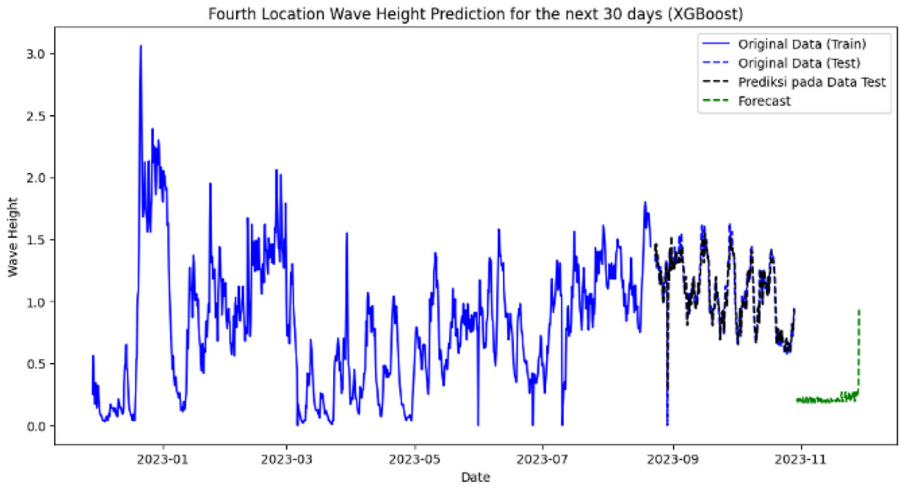

Fig. 18. Next 30-day wave height prediction.

Table 3
Method comparison results.

Location	Coordinate	LSTM			XGBoost		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE
1st Location	−6.21;111.80	**0,045**	**0,029**	**8671 %**	0,046	0,031	9022 %
2nd Location	−6.22;112.40	0,051	**0,035**	**10,64 %**	**0,050**	0,036	11,05 %
3rd Location	−6.26;112.20	0,044	**0,027**	**7773 %**	**0,043**	0,028	8160 %
4th Location	−6.31;112.27	0,042	0,025	7386 %	**0,040**	**0,025**	**7286 %**

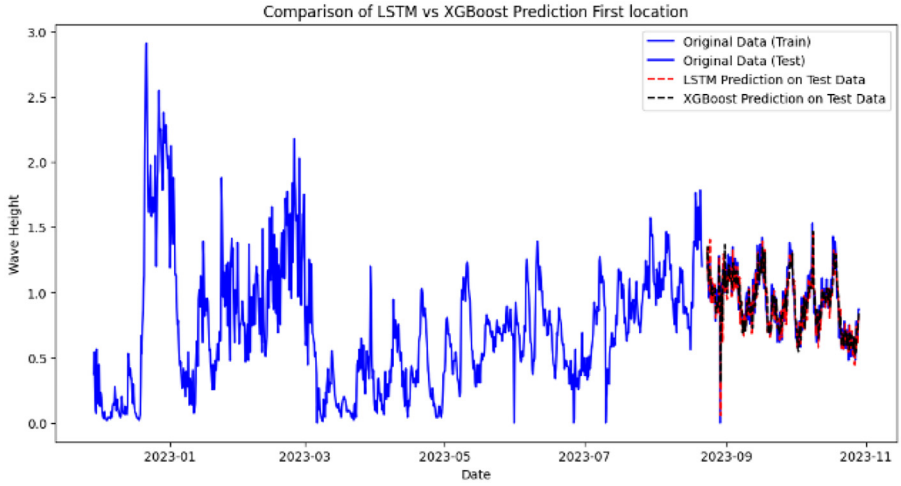

Fig. 19. Comparison of LSTM and XGBoost at location 1.

Based on Table 3, the results of the comparison of the two models in predicting wave heights at the first location show that the LSTM method is more optimal when compared to the XGBoost method, the RMSE value of the LSTM model is 0.045, MAE is 0.029 and MAPE is 8.671 %, but the difference between the two methods is relatively small. Fig. 19 shows the results of the comparison plot of the two methods, it can be seen that the prediction plot of the LSTM method has a value that tends to be lower when compared to the prediction results with the XGBoost method. However, the LSTM and XGBoost methods have fairly good accuracy results, this is known from the plots of the two methods which have almost the same pattern as the actual data plot Fig. 20.

The results of the comparison of the two models for the second location also have relatively small difference values for the RMSE, MAE, and MAPE values for the two methods of 0.001; 0.001; and 0.41, respectively. The predicted value of the LSTM method has MAE and MAPE values that tend to be better when compared to the XGBoost method, namely 0.035 and 10.64 %. The XGBoost method has a better RMSE value when compared to LSTM which is 0.050. Fig. 21 shows the results of the comparison plot of the two

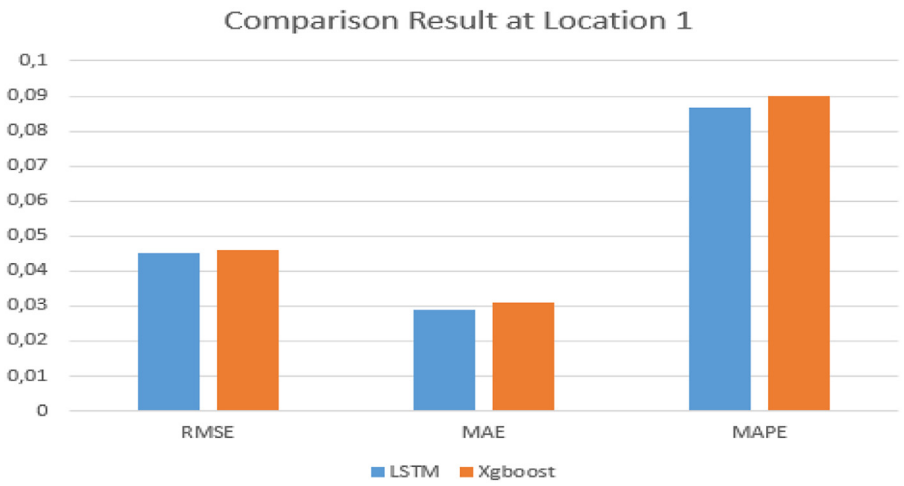

Fig. 20. Bar chart of result LSTM and XGBoost at location 1.

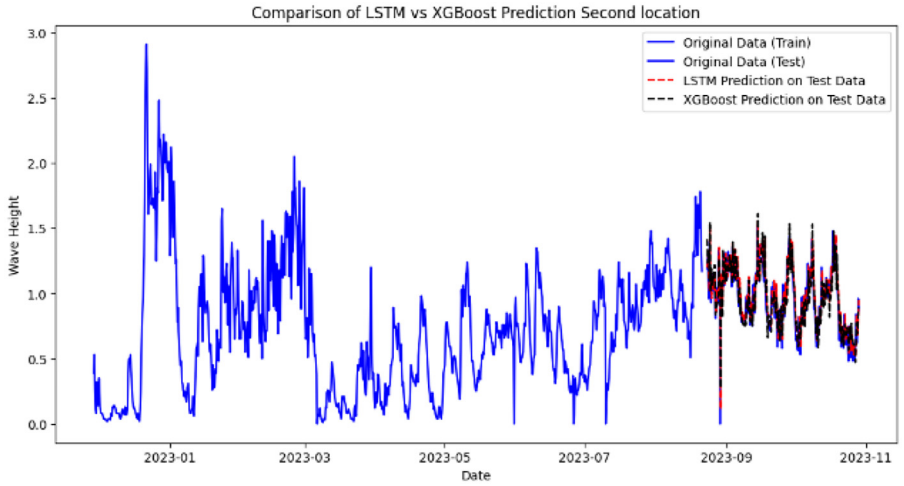

Fig. 21. Comparison of LSTM and XGBoost at location 2.

methods, It can be seen that the prediction plot of the XGBoost method has a value that tends to be higher when compared to the prediction results with the LSTM method. However, the LSTM and XGBoost methods have fairly good accuracy results, this is known from the plots of the two methods which have a pattern that is quite similar to the actual data plot of the second location Fig. 22.

The evaluation results of the LSTM method at the third location have a better value when compared to the XGBoost method. However, the difference between the evaluation results of the two methods is relatively small. The difference for RMSE, MAE, and MAPE values for both methods are 0.001; 0.001; and 0.387. From Fig. 23, the XGBoost prediction results show higher results in early September 2023 when compared to the actual data and LSTM prediction plots. However, from the results of the comparison plot of the two methods, it can be seen that both method plots have almost the same pattern and height Fig. 24.

The results of the comparison of the two models for the fourth location obtained the same MAE value between the two methods of 0.25, while the RMSE and MAPE values between the two methods have relatively small differences. The difference between the RMSE and MAPE values for the two methods is 0.002 and 0.100, respectively. The LSTM and XGBoost methods have quite good accuracy results, this is known from the plots of both methods (Fig. 25) which have almost the same pattern as the actual data plot. In addition, the plot results of both methods at the fourth location also show results that are more fitted to the actual data when compared to plots from the first, second, and third locations Fig. 26.

Time efficiency of LSTM and XGBoost

The research process using the Long Short-Term Memory (LSTM) method ran smoothly, with the optimal parameters for each location determined using grid search, resulting in optimal predictions on the test data. This is further supported by early stopping to prevent overfitting and underfitting. The computation time for parameter determination for the LSTM method was 35 min for each

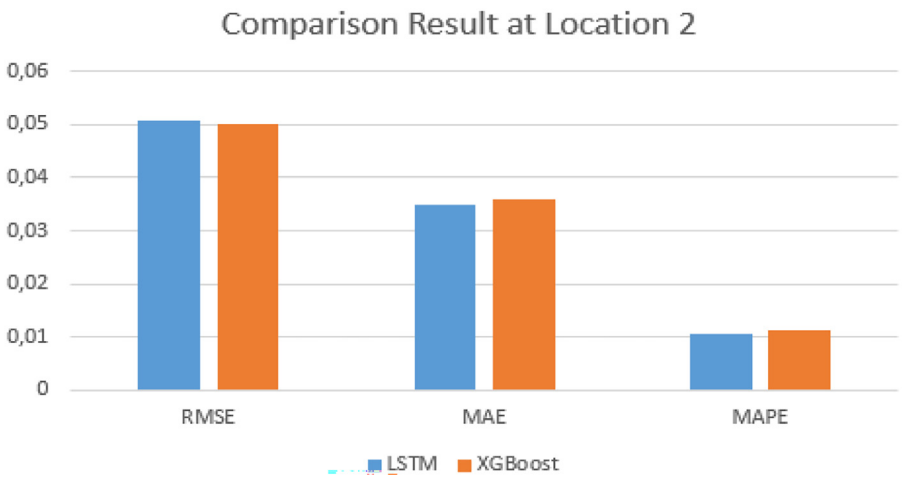

Fig. 22. Bar chart of result LSTM and XGBoost at location 2.

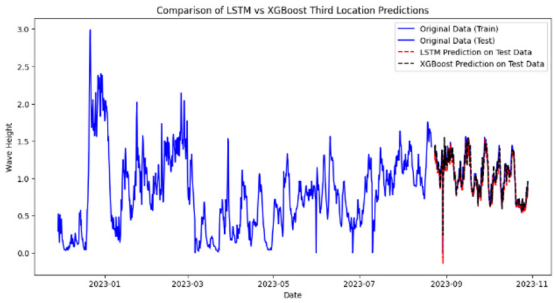

Fig. 23. Comparison of LSTM and XGBoost at location 3.

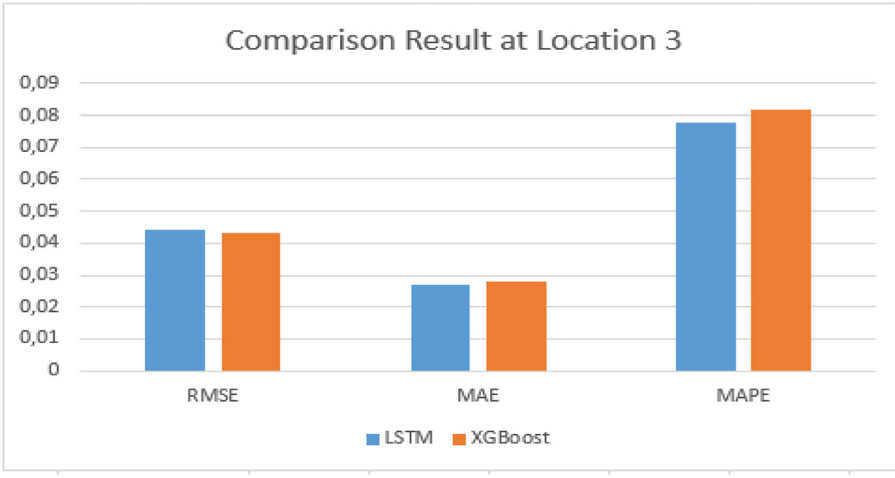

Fig. 24. Bar chart of result LSTM and XGBoost at location 3.

location. The LSTM method produced predictions at the first location (RMSE 0.045; MAE 0.029; MAPE 8.671 %), second location (RMSE 0.051; MAE 0.035; MAPE 10.64 %), third location (RMSE 0.044; MAE 0.027; MAPE 7.773 %), and fourth location (RMSE 0.042; MAE 0.025; MAPE 7.386 %). The research process using the Extreme Gradient Boosting (XGBoost) method also ran smoothly, with the optimal parameters for each location determined using grid search, resulting in optimal predictions on the test data. The use of L2 regularization helped prevent overfitting and underfitting. The computation time for parameter determination for the XGBoost method was 5 min for each location. The XGBoost method produced predictions at the first location (RMSE 0.046; MAE 0.031; MAPE

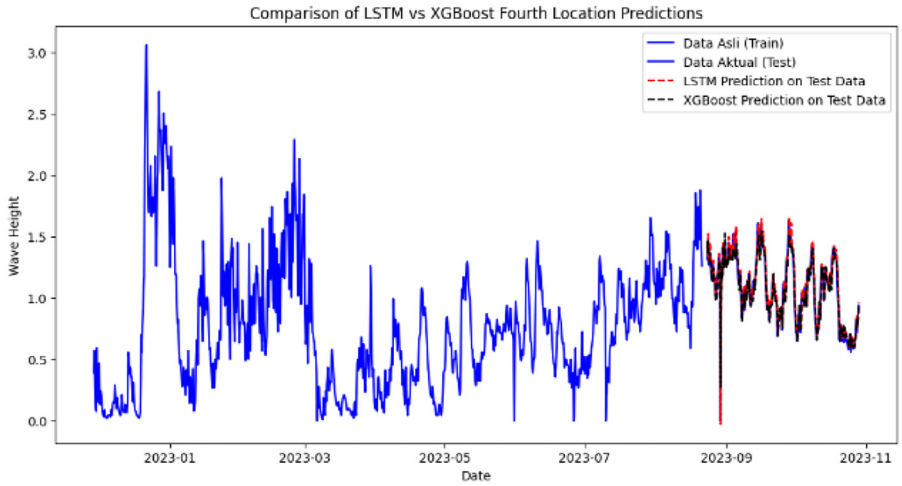

Fig. 25. Comparison of LSTM and XGBoost at location 4.

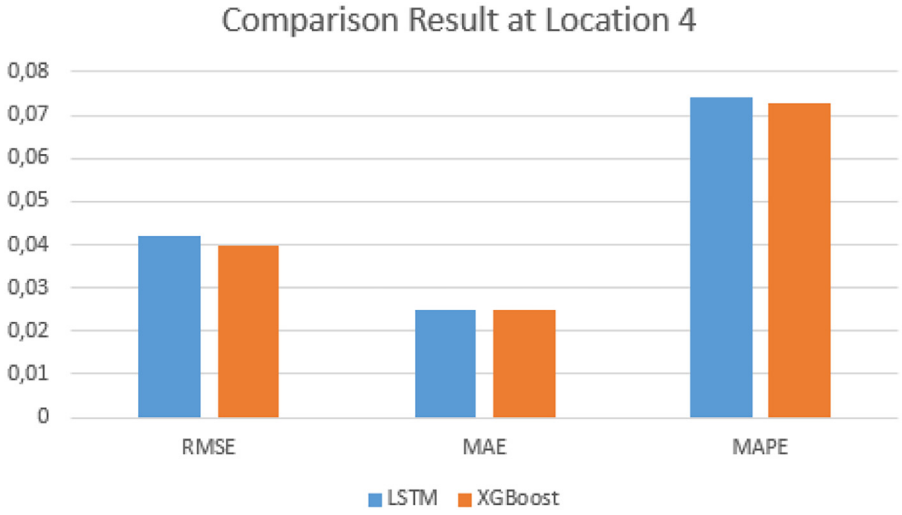

Fig. 26. Bar chart of result LSTM and XGBoost at location 4.

9.022 %), second location (RMSE 0.050; MAE 0.036; MAPE 11.05 %), third location (RMSE 0.043; MAE 0.028; MAPE 8.160 %), and fourth location (RMSE 0.040; MAE 0.025; MAPE 7.286 %).

Based on the research results, several conclusions can be drawn regarding comparing Long Short-Term Memory (LSTM) and Extreme Gradient Boosting (XGBoost) methods for predicting sea wave heights in Tuban Regency to ensure fishermen's safety. The LSTM method proves to be superior in predicting wave heights at the first location (RMSE 0.045; MAE 0.029; MAPE 8.671 %), second location (RMSE 0.051; MAE 0.035; MAPE 10.64 %), and third location (RMSE 0.044; MAE 0.027; MAPE 7.773 %). In contrast, XGBoost performed better at the fourth location (RMSE 0.040; MAE 0.025; MAPE 7.286 %). Despite the relatively small differences, both methods demonstrate optimal performance for predicting sea wave heights in Tuban Regency. The predictions for the next 30 days indicate that wave heights across all locations remain below the safety limit set by BMKG, which is <1.25 m, allowing fishermen in Tuban Regency to safely continue their activities throughout most of the year. However, it is crucial to note that from December to March, predictions indicate wave heights may exceed 1.25 m, which could pose risks for fishing activities during that period.

Limitations

Although this research contributes to understanding Comparative analysis of Long Short-Term Memory and Extreme Gradient Boosting methods to predict the ocean wave in Tuban Regency for fisherman safety, there are some limitations that future studies may address :

1. Single Factor Analysis: The study focused exclusively on wave height as the only variable, which may limit the comprehensiveness of the analysis. Future research should consider integrating additional environmental factors impacting wave predictions and fishermen's safety.
2. Geographical Context: The applicability of the models may be limited to the specific geographical conditions of Tuban Regency. Future researchers can involve spatial data such as seabed topography, including depth and sea slope. Integration of these data can provide a better understanding of the physical conditions of the region and increase the number of features that can be used by the model.
3. External Influences: The research did not account for external factors such as operational constraints, and equipment-specific variables, which could influence the accuracy of the predictions. Addressing these factors in future research could enhance the robustness and practical utility of the models.

Ethics statements

As an expert scientist and along with co-authors of the concerned field, the paper has been submitted with full responsibility, following due ethical procedure, and there is no duplicate publication, fraud, plagiarism, or concerns about animal or human experimentation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Riswanda Ayu Dhiya'ulhaq: Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Software. **Anisya Safira:** Conceptualization, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Software. **Indah Fahmiyah:** Supervision, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Validation. **Mohammad Ghani:** Supervision, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Validation.

Data availability

Data will be made available on request.

Acknowledgments

The data source used in this research is secondary data obtained from Perak Meteorological Station Surabaya II.

References

[1] ILO, Fishers first - good practices to end labour exploitation at sea, in: International Labour Organization: Fundamental Principals and Rights at Work Branch (FUNDAMENTALS), Sectoral. Policies Department (SECTOR), Geneva, 2016, p. 82.
[2] A. Husnah, A. Abdillah, Vera Mandailina, S. Syaharudin, S. Mehmood, Wind speed regression model in forecasting wave height in the shipping channel zone, JST (J. Sains dan Teknol. 12 (1) (2023) 30–38, doi:10.23887/jstundiksha.v12i1.50981.
[3] D. Adytia, A.R. Yonanta, N. Subasita, Wind wave prediction by using autoregressive integrated moving average model : case study in Jakarta Bay, Int. J. Inf. Commun. Technol. 4 (2) (2019) 33, doi:10.21108/ijoict.2018.42.300.
[4] J. Liu, Navigating the financial landscape: the power and limitations of the ARIMA model, Highlights Sci. Eng. Technol. 88 (2024) 747–752, doi:10.54097/9zf6kd91.
[5] R.M. Campos, M.O. Costa, F. Almeida, C.G. Soares, Operational wave forecast selection in the Atlantic ocean using random forests, J. Mar. Sci. Eng. 9 (3) (2021), doi:10.3390/jmse9030298.
[6] S.N. Londhe, V. Panchang, One-day wave forecasts based on artificial neural networks, J. Atmos. Ocean. Technol. 23 (11) (2006) 1593–1603, doi:10.1175/JTECH1932.1.
[7] F.C. Minuzzi, L. Farina, A deep learning approach to predict significant wave height using long short-term memory, Ocean Model. 181 (December) (2023), doi:10.1016/j.ocemod.2022.102151.
[8] F.A.R. Abdullah, N.S. Ningsih, T.M. Al-Khan, Significant wave height forecasting using long short-term memory neural network in Indonesian waters, J. Ocean Eng. Mar. Energy 8 (2) (2022) 183–192, doi:10.1007/s40722-022-00224-3.
[9] J. Zhang, et al., Improving wave height prediction accuracy with deep learning, Ocean Model, 188, 2024, doi:10.1016/j.ocemod.2023.102312.
[10] Z.F. Meng, Z. Chen, B.C. Khoo, A.M. Zhang, Long-time prediction of sea wave trains by LSTM machine learning method, Ocean Eng 262 (August) (2022) 112213, doi:10.1016/j.oceaneng.2022.112213.
[11] D.D. Pramesti, D.C.R. Novitasari, F. Setiawan, H. Khaulasari, Long-Short Term Memory (LSTM) for predicting velocity and direction sea surface current on Bali strait, BAREKENG J. Ilmu Mat. dan Terap. 16 (2) (2022) 451–462, doi:10.30598/barekengvol16iss2pp451-462.
[12] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceeding of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17, 2016, pp. 785–794, doi:10.1145/2939672.2939785.
[13] H. Hu, A.J. van der Westhuysen, P. Chu, and A. Fujisaki-Manome, "Predicting Lake Erie wave heights and periods using XGBoost and LSTM," Ocean Model., vol. 164, no. November 2020, p. 101832, 2021, doi: 10.1016/j.ocemod.2021.101832.
[14] F. Anggraeni, D. Adytia, A.W. Ramadhan, Forecasting of wave height time series using AdaBoost and XGBoost, case study in Pangandaran, Indonesia, in: 2021 International Conference on Data Science and Its Applications (ICoDSA), IEEE, Oct. 2021, pp. 97–101, doi:10.1109/ICoDSA53588.2021.9617524.
[15] D. Tarwidi, S.R. Pudjaprasetya, D. Adytia, M. Apri, An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach, MethodsX 10 (December 2022) (2023) 102119, doi:10.1016/j.mex.2023.102119.

[16] A. Callens, D. Morichon, S. Abadie, M. Delpey, B. Liquet, Using random forest and gradient boosting trees to improve wave forecast at a specific location, Appl. Ocean Res. 104 (August) (2020) 102339, doi:10.1016/j.apor.2020.102339.
[17] Eni, Proceeding the 1st international seminar on sustainability in the marine fisheries sector 2017.establishing sustainable marine and fisheries sector to support food security within ASEAN economic community framework, Angew. Chem. Int. Ed. 6 (11) (2017) 951–952 no. Mi, pp. 5–24.
[18] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, doi:10.1162/neco.1997.9.8.1735.
[19] K. Prakash, R. Kannan, A.S.A, and K.G.R, Advanced deep learning for engineers and scientists. 2021. [Online].Available: https://books.google.com.pk/books?id=MDsNzgEACAAJ%0Ahttps://link.springer.com/10.1007/978-3-030-66519-7.
[20] R.C. Staudemeyer and E.R. Morris, "Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks," pp. 1–42, 2019, [Online]. Available: http://arxiv.org/abs/1909.09586.
[21] A. Ghosh, S. Bose, G. Maji, N.C. Debnath, S. Sen, Stock price prediction using LSTM on Indian share market, Epic Ser. Comput. 63 (September 2019) (2019) 101–110, doi:10.29007/qgcz.
[22] P. Marco, Time Series Forecasting in Python, Simon and Schuster, New York, 2022.
[23] D.M. Belete, M.D. Huchaiah, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results, Int. J. Comput. Appl. 44 (9) (2022) 875–886, doi:10.1080/1206212X.2021.1974663.
[24] D.A. Anggoro, S.S. Mukti, Performance comparison of grid search and random search methods for hyperparameter tuning in extreme gradient boosting algorithm to predict chronic kidney failure, Int. J. Intell. Eng. Syst. 14 (6) (2021) 198–207, doi:10.22266/ijies2021.1231.19.
[25] K. Vijayaprabakaran, K. Sathiyamurthy, Towards activation function search for long short-term model network: a differential evolution based approach, J. King Saud Univ. - Comput. Inf. Sci. 34 (6) (2022) 2637–2650, doi:10.1016/j.jksuci.2020.04.015.
[26] E. Ismanto, N. Effendi, An LSTM-based prediction model for gradient-descending optimization in virtual learning environments, Comput. Sci. Inf. Technol. 4 (3) (2023) 199–207, doi:10.11591/csit.v4i3.p199-207.
[27] S. Malik, R. Harode, and A. Singh Kunwar, "XGBoost: a deep dive into boosting Feb 3 · 12 min read," no. February 2020, doi: 10.13140/RG.2.2.15243.64803.
[28] M.M.K. Kazemi, Z. Nabavi, M. Khandelwal, Prediction of blast-induced air overpressure using a hybrid machine learning model and gene expression programming (GEP): a case study from an iron ore mine, AIMS Geosci. 9 (2) (2023) 357–381, doi:10.3934/geosci.2023019.
[29] S.V. Murty, R.Kiran Kumar, Accurate liver disease prediction with extreme gradient boosting, Int. J. Eng. Adv. Technol. 8 (6) (2019) 2288–2295, doi:10.35940/ijeat.F8684.088619.
[30] A.M. Mohammad Mirzehi Kalateh Kazemi, Zohreh Nabavi, Mojtaba Rezakhah, Application of XGB-based metaheuristic techniques for prediction time-to-failure of mining machinery, Syst. Soft Comput. 5 (2023) [Online]. Available https://www.sciencedirect.com/science/article/pii/S2772941923000145 .
[31] Z. Nabavi, M. Mirzehi, H. Dehghani, P. Ashtari, A hybrid model for back-break prediction using XGBoost machine learning and metaheuristic algorithms in chadormalu iron mine, J. Min. Environ. 14 (2) (2023) 689–712, doi:10.22044/jme.2023.12796.2323.
[32] E. Ding, Regularization, Data Sci. Resour. (2022) 1–8 [Online]. Available https://s3.amazonaws.com/kajabi-storefronts-production/file-uploads/sites/2147512189/themes/2150624317/downloads/5355dc-82-8b2-2b53-dd6a85bcb82_Regularization.pdf .
[33] L.B.V. de Amorim, G.D.C. Cavalcanti, R.M.O. Cruz, The choice of scaling technique matters for classification performance, Appl. Soft Comput. 133 (2023) 1–37, doi:10.1016/j.asoc.2022.109924.
[34] H.K. Jabbar and R.Z. Khan, "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)," no. December 2014, pp. 163–172, 2015, doi: 10.3850/978-981-09-5247-1_017.
[35] F. Lazzeri, Machine learning for time series forecasting with python. 2020. doi: 10.1002/9781119682394.
[36] H. Hewamalage, K. Ackermann, C. Bergmeir, Forecast evaluation for data scientists: common pitfalls and best practices, Data Min. Knowl. Discov. 37 (2) (2023) 788–832, doi:10.1007/s10618-022-00894-5.
[37] Z. Chang, Y. Zhang, W. Chen, Effective Adam-optimized LSTM neural network for electricity price forecasting, IEEE Int. Conf. Softw. Eng. Serv. Sci. 00 (Figure 1) (2018) 245–248.
[38] N.M.R. Putri Udiani, I.K.G. Darma Putra, G.M. Arya Sasmita, Forecasting of Arabica coffee production in Bali province using support vector regression, Int. J. Comput. Appl. Technol. Res. 9 (2) (2020) 41–46, doi:10.7753/ijcatr0902.1001.