An ensemble model with feature selection for nearshore wave forecasting

Chu Thi Quyen*, Ngo Thi Thanh Hoa, Nguyen Thi Cam Ngoan

Hanoi University of Industry, 298 Cau Dien, Tay Tuu, Hanoi, Vietnam.
*Corresponding author: chuthiquyen_cntt@haui.edu.vn
Received 28 Mar. 2025; Revised 6 May 2025; Accepted 10 Aug. 2025; Published 25 Aug. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.105.2025.121-129

ABSTRACT

The study proposes an ensemble one-week ahead Wave Forecast of Nearshore Waves (OWFNW) framework for managing shipping and construction in marine work sites. The framework uses XGBoost with feature selection (FS_XGBoost) for forecasting at 5 stations on the Japanese coast. XGBoost-based wave models are developed for each station, transforming global wave data into nearshore wave predictions. Models are trained using four different training sets from the Japan Meteorological Agency (JMA), National Oceanic and Atmospheric Administration (NOAA), European Centre for Medium-Range Weather Forecasts (ECMWF) and Nationwide Ocean Wave information network for Ports and HarbourS (NOWPHAS). The results indicate that selecting features enhances the model's prediction accuracy and refining prediction errors. The methodology can be applied to other nearshore seas where global wave forecast data is available.

Keywords: Feature selection; XGBoost; Nearshore wave; Global wave forecast.

1. INTRODUCTION

Accurate long-range forecasting of nearshore wave conditions up to a week in advance is a significant challenge for coastal infrastructure planning, disaster preparedness, and maritime safety. While traditional numerical models are computationally intensive, Machine Learning (ML) and Artificial Intelligence (AI) offer data-driven alternatives [1 - 4]. However, the effectiveness of ML models is heavily reliant on the selection of relevant input features to avoid issues like noise, overfitting, and increased computational cost. This study puts forward an ensemble learning framework that uses optimized feature selection to forecast nearshore waves up to 168 hours ahead, aiming to improve upon traditional ML techniques.

The rest of this paper is structured as follows: The machine learning models and datasets utilized in this investigation are described in section 2. The process for converting global wave data into nearshore predictions is described in section 3. The outcomes and performance assessment of the models are shown in section 4. Section 5 concludes by discussing the findings' ramifications and possible applicability to other nearshore environments.

2. RELATED WORKS

ML and AI in wave forecasting

The application of ML and AI techniques in wave forecasting has gained significant attention in recent years due to their ability to model nonlinear oceanic processes. Early studies employed traditional ML models such as Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) to predict wave heights and periods using historical data and meteorological inputs. More recent approaches have leveraged deep learning models, including Convolutional Neural Networks (CNNs) [7, 8] and Long Short-Term Memory (LSTM) networks [9, 10], to capture spatial and temporal dependencies in wave dynamics. For instance, [11] developed an LSTM-based model that demonstrated superior predictive accuracy compared to conventional numerical models. Similarly, [12] explored hybrid deep learning frameworks that integrate physics-based constraints with AI-driven forecasting.

Feature selection for wave forecasting

The selection of input variables plays a crucial role in determining the accuracy and efficiency of ML-based wave forecasting models. Numerous studies have explored optimal feature selection techniques to enhance forecasting performance while minimizing computational load. For instance, the approach introduced in [13] employed a Bidirectional Long Short-Term Memory (BiLSTM) network, leveraging highly spatially correlated wind data to identify the most relevant features for wave forecasting. This method demonstrated superior predictive accuracy compared to conventional models such as LSTM, Support Vector Regression (SVR), and Generalized Regression Neural Networks (GRNN). Similarly, study [14] presented a hybrid feature selection approach based on imbalance learning (HFS-IL), which integrates an LSTM-based imbalance discriminator with a hybrid selection algorithm to determine optimal input subsets. The results showed that HFS-IL effectively mitigates data imbalance issues and enhances prediction accuracy. Extending the findings of Kim [9], who demonstrated that GMDH and ANN yielded the best performance using NOAA and ECMWF data for wave height, and JMA and ECMWF data for wave period at the Port of Hitachinaka, the current study proposes an embedded XGBoost framework to automate and optimize feature selection.

Ensemble learning for wave prediction

Ensemble learning is widely recognized for enhancing the robustness and generalization of ML models in wave forecasting. Techniques such as random forests, boosting algorithms (e.g., XGBoost, LightGBM), and stacked models have shown strong performance. For example, [15] proposed an Ens-ELM model for daily wave height prediction across multiple regions, outperforming ELM, OS-ELM, and SVR. Study [16] introduced a stacked ensemble model that achieved high accuracy for monthly wave heights, while [17] improved lightweight models by updating forecasts based on past errors. Despite progress, long-term nearshore forecasts remain difficult. This study proposes an ensemble model with optimized feature selection, building on prior findings that emphasize the importance of tailored input and model design.

3. MATERIALS AND METHOD

3.1. Global wave data and nearshore wave data

This study employs global wave forecasts from ECMWF, JMA, and NOAA, each offering distinct model configurations. ECMWF's HRES-WAM is coupled with an atmospheric model at 0.125° resolution, while JMA's GWM and NOAA's WW3 are standalone models at 0.5° resolution. All models support 36 directions and differ in frequency resolution and forecast settings. Forecast data (wave height and period) at 12 UTC for seven days ahead were collected at five sites, along with observed data from NOWPHAS. An XGBoost-based framework, as figure 1, is used to transform global wave inputs into nearshore predictions. For each station, separate models are trained to forecast significant wave height (Hp) and peak period (Ts), tailored to local conditions.

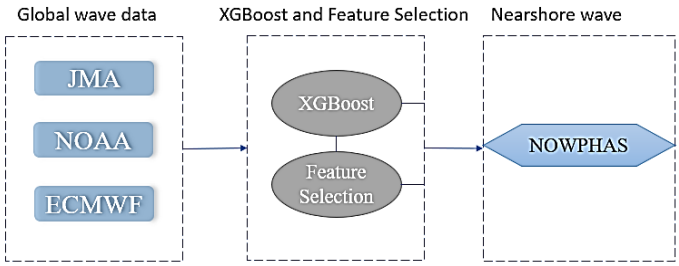

Figure 1. *The proposed framework: transforming global wave forecasts into nearshore wave predictions using station-specific XGBoost models.*

3.2. A framework of XGBoost-based wave forecast

3.2.1. Embedded feature selection

Feature selection techniques can be divided into three categories: filter methods, wrapper methods, and embedded methods, of which embedded feature selection is an advanced technique that integrates feature selection within the model training process. It leverages algorithms that inherently evaluate feature importance, such as decision trees, Lasso regression, and boosting methods like XGBoost. This approach balances predictive performance and computational efficiency by selecting the most relevant features while training the model. As a result, embedded methods often outperform filter and wrapper approaches in handling high-dimensional datasets.

3.2.2. Analysis the input features importance

In XGBoost, the weight method calculates feature importance based on the frequency of a feature being used to split the data across all trees in the model. Specifically, it counts how many times a feature is used to make a split in the decision trees. Features that are used more frequently are considered more important.

Algorithm to compute feature importance using weight:

- Build trees: During the training process, XGBoost constructs multiple decision trees.
- Count Splits: For each feature, count the number of times it is used to split the data across all trees.
- Normalize (Optional): The counts can be normalized to represent the relative importance of each feature.
- Mathematically, the importance of a feature f is computed as:

$$i_f = \sum_{t=1}^{N} \mathcal{L}(f \; is \; used \; in \; tree \; t)$$

Where N is the total number of trees, $\mathcal{L}$ is an indicator function that returns 1 if the feature f is used in tree t, and 0 otherwise.

3.2.3. XGBoost model with feature selection

This study employed a hybrid embedded feature selection method using XGBoost called FS_XGBoost, which retains only the most important features based on the feature importance scores from a trained model. It evaluates each feature's importance against a specified threshold and eliminates those deemed less significant. The algorithm for this approach is as follows:

Algorithm FS_XGBoost(X, y, thresholds):

Initialize best_accuracy ← 0

Initialize optimal_num_features ← 0

For each threshold in thresholds:

1. Train an initial XGBoost model on (X, y)
2. Obtain feature importances from the trained model
3. Select features whose importance ≥ threshold
4. Create a new dataset X_selected using selected features
5. Split X_selected and y into training and test sets
6. Train a new XGBoost model on the training set
7. Evaluate the model on the test set to get accuracy
8. If accuracy > best_accuracy:

a. Update best_accuracy

b. Update optimal_num_features ← number of selected features

Return optimal_num_features

4. EXPERIMENTS AND RESULTS

4.1. Experiments

The dataset used in this study consists of wave height and period measurements from five coastal locations in Japan: Onahama, Akita, Kanazawa, Hososhima, and Naha. They were collected in 2019 and 2020. After removing missing records and normalizing by the standard normal distribution, we divided them into training and testing data, as shown in table 1. The data were taken at the stations in real-time, 1 day ahead, 3 days ahead, 5 days ahead, and 7 days ahead of 3 observatories: JMA, NOAA, and ECMWF. The input data of the model includes InitJMA, 24hJMA, 72hJMA, 120hJMA, and 168hJMA, which correspond to the current time, one day ago, three days ago, and seven days ago of JMA data; InitECMWF, 24hECMWF, 72hECMWF, 120hECMWF, and 168hECMWF for the corresponding ECMWF data; and InitNOAA, 24hNOAA, 72hNOAA, 120hNOAA, and 168hNOAA for the NOAA data. The output data from NOWPHAS is significant nearshore wave height, H_p, or period, T_s, observed data *from* NOWPHAS.

Table 1. *The data size at the stations.*

		H_p		T_S	
No	**Location**	**Training set**	**Testing set**	**Training set**	**Testing set**
1	Onahama	423	206	414	207
2	Naha	389	194	401	185
3	Akita	396	198	397	198
4	Kanazawa	138	71	139	70
5	Hososhima	381	190	381	191

We will briefly explain five other machine learning methods tried out in our experiments. They are the Artificial Neural Network (ANN), Random Forest (RF), Support Vector Regression (SVR), Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). Whereas ANN is used in many wave forecasting problems, Random Forest is favored for its robustness, handling of non-linear relationships, and feature importance insights. SVR is selected for its effectiveness in modeling complex, non-linear data using kernel functions and margin optimization; CNN is used for regression tasks involving spatial data, as it can automatically extract important features and capture local patterns efficiently; and LSTM excels at modeling long-range patterns in sequences via gated memory control. The parameters of the methods are given in table 2.

Table 2. *The parameters of ML models.*

SVM	SVM Type: epsilon-SVM C: 7 Epsilon: 0.001	XGBoost	learning_rate: 0.01 max_depth: 5 n_estimators: 400 eval_metric 'rmse'
RF	n_estimators: 50	ANN (MLP)	Hidden layers 3 (hidden_layer_sizes (128, 64, 32)) learning rate: 'adaptive' activation: ‘relu’ solver: ‘adam’ max_iter: 10000 batch_size: 32
CNN		Conv1D_1: Conv1D Filters = 64, Kernel size = 3, Activation = ReLU Dropout_1: Dropout Dropout rate = 0.5 Conv1D_2: Conv1D Filters = 32, Kernel size = 3, Activation = ReLU	

	Dropout_2: Dropout Dropout rate = 0.4 Flatten : Flattens feature maps into 1D Dense_1: Fully Connected Units = 32, Activation = ReLU Output: Dense Units = 1, Activation = Linear (regression output) Optimizer: Adam Loss Function: Mean Squared Error (MSE) Epochs:30 Batch Size: 50
LSTM	LSTM_1: LSTM (64, ReLU) LSTM_2: LSTM (32, ReLU) Dropout_1: Dropout rate = 0.5 Dense_1: Fully Connected 16 ReLU Dropout_2: Dropout rate = 0.4 Dense_2: Fully Connected 8 ReLU Output: Fully Connected 1 Linear Regression output Optimizer: Adam Loss Function: Mean Squared Error (MSE) Epochs:50 Batch Size: 32

The model performance reflects the accuracy of the model. The Normalize Root Mean-Squared Error (NRMSE) and the Coefficient of Correlation (CC) are the metrics used in this study to assess the performance. The statistical equations for these metrics are given in equations (1), (2), where $y_{pre,i}$ is the predicted value obtained with the model and $y_{obs,i}$ is the measured value from the stations. $\bar{y}_{obs}$ is the average of the actual values, $\bar{y}_{pre}$ is the average of the predicted values for the entire dataset. $y_{obs,max}$, $y_{obs,min}$ are respectively the maximum and minimum values in the entire data set.

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{obs,i} - y_{pre,i})}}{(y_{obs,max} - y_{obs,min})} \quad (1)$$

$$CC = \frac{\sum_{i=1}^{n}(y_{obs,i} - \bar{y}_{obs})(y_{pre,i} - \bar{y}_{pre})}{\sqrt{\sum_{i=1}^{n}(y_{obs,i} - \bar{y}_{obs})^2}\sqrt{\sum_{i=1}^{n}(y_{pre,i} - \bar{y}_{pre})^2}} \quad (2)$$

4.2. Results and discussion

4.2.1. Features importance analysis

The feature importance rankings for the XGBoost models show that the top five contributors for both wave height (Hp) and wave period (Ts) predictions are consistently the ECMWF forecasts at 168 h, 120 h, 72 h, 24 h, and initialization time. The 168hECMWF feature is the most influential, contributing over 17.5% for Hp, and exceeding 30% at Onahama. For Ts, feature contributions are more evenly distributed, with 168hECMWF still leading at around 12%, and over 20% at Akita. In contrast, JMA data features contribute less than 5% across all models. Additionally, only two optimal Hp models included JMA features, while none were used in the optimal Ts models.

4.2.2. Model performance

a. Comparing to other machine learning algorithms

We conduct experiments to compare our approach with machine learning and deep learning

techniques used in wave forecasting. We used NRMSE and CC values mentioned in formulas 1 and 2 above that are crucial for assessing the model's fit to the raw data. After conducting a feature importance analysis, input features were chosen based on their scores to construct various XGBoost models and identify the best one. It can be observed from tables 3 and 4 that for the H_p data, the CC values for the testing data range of FS_XGBoost from 0.8 to 0.95, for the optimal number of features selected. It is observed that among the models tested, FS_XGBoost consistently achieved the highest correlation coefficients and lowest errors.

b. Performance of feature selection

Tables 3 to 5 presents the selected features and accuracy for both the XGBoost with feature selection and *the* original XGBoost methods, which utilize all available features. Table 5 reveals that the model trained with the feature set automatically identified by the proposed approach outperforms all baseline models for both H_p and T_s. This indicates that feature selection through the method FS_XGBoost is more effective than relying solely on traditional algorithms.

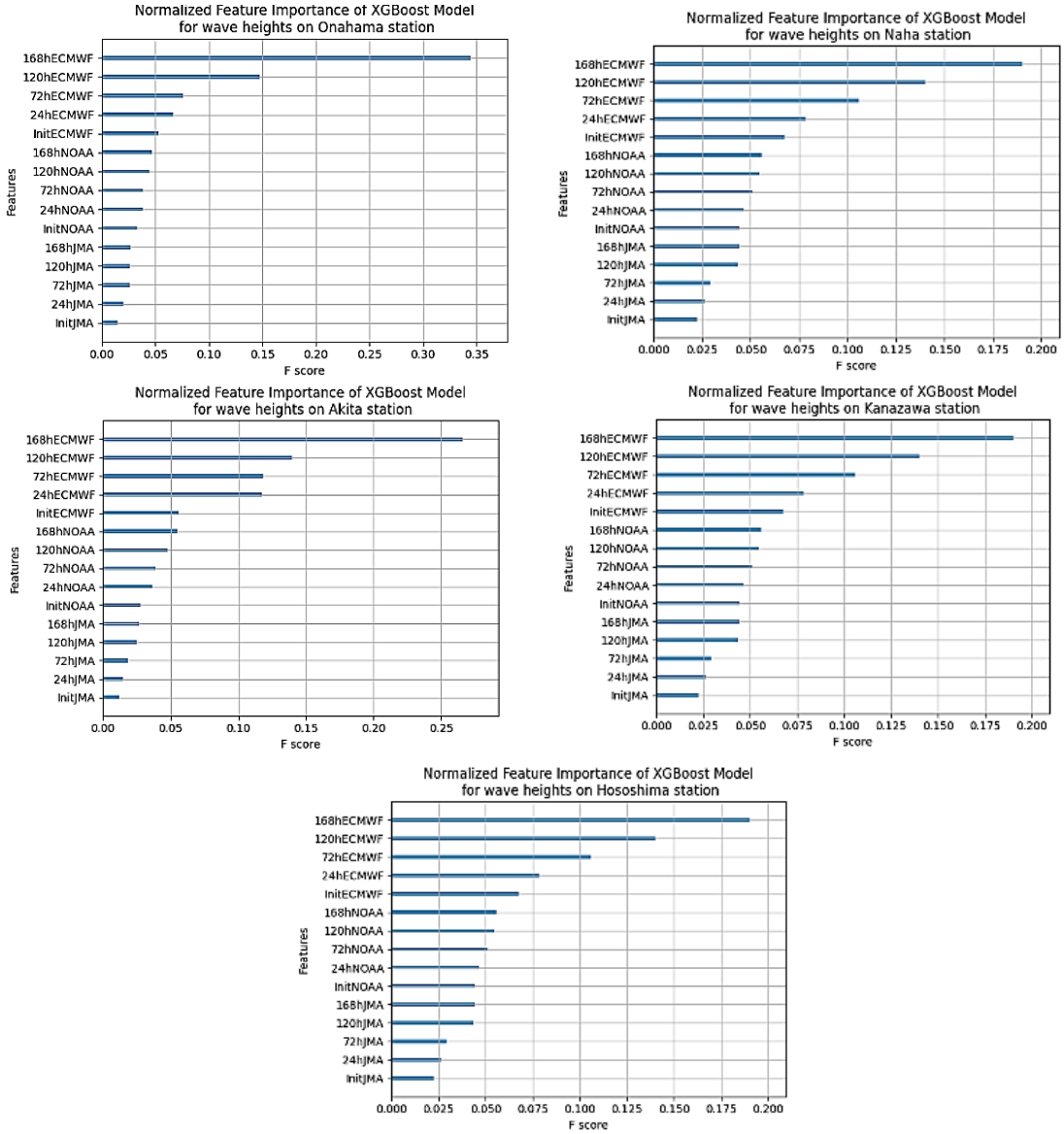

Figure 2. *Feature importance detailed scores of the H_p model.*

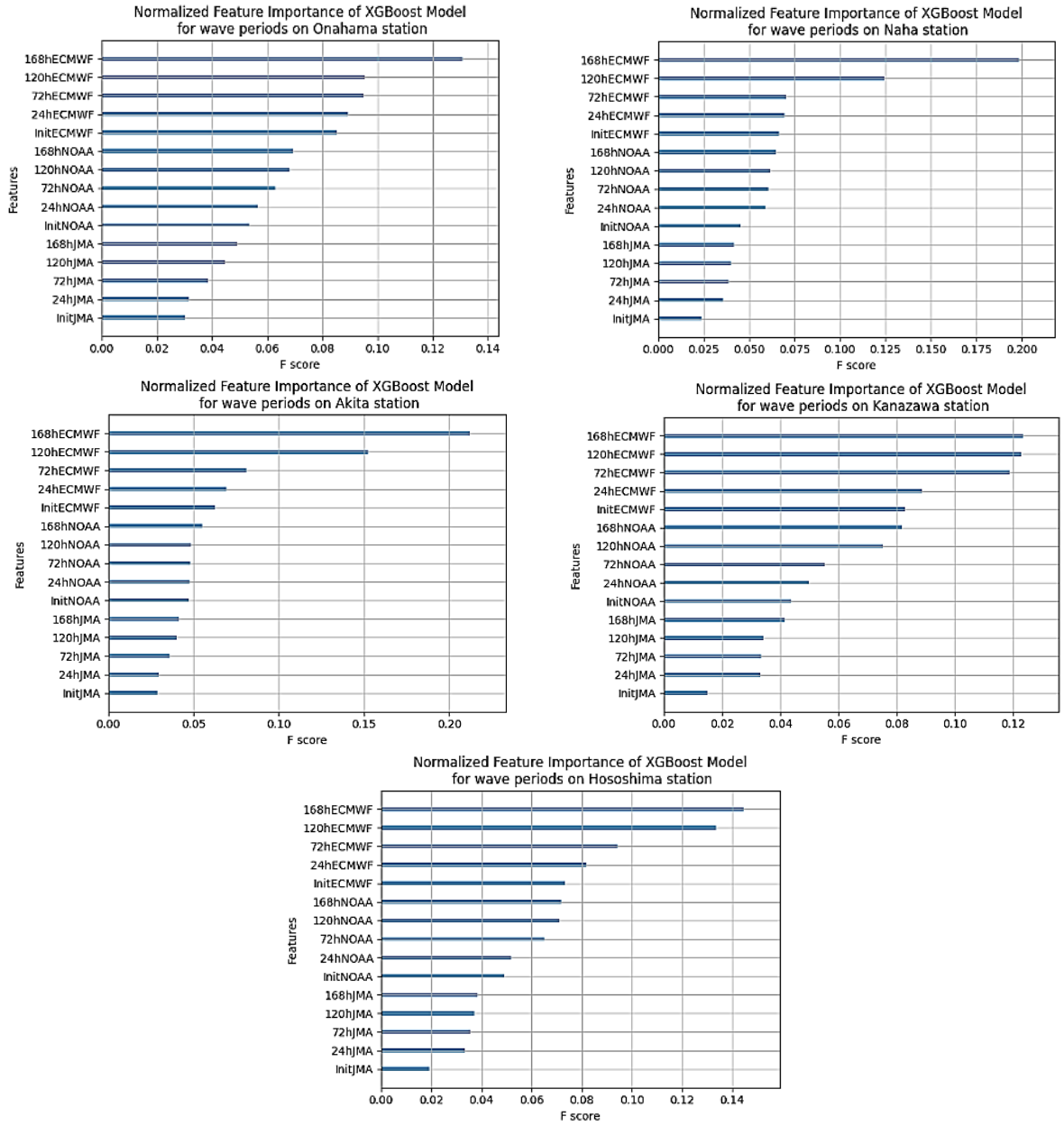

Figure 3. *Feature importance detailed scores of the T_S model.*

Table 3. *The indices of waves forecasted by the models for H_p.*

Station	FS_XGBoost		CNN		LSTM		SVM		ANN		RF	
	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)
Onahama	**0.943**	**5.54**	0.934	7.38	0.938	6.35	0.938	6.27	0.930	5.99	0.942	5.58
Kuzi	**0.951**	**6.64**	0.941	8.04	0.940	7.97	0.948	7.18	0.936	7.55	0.948	6.86
Akita	**0.933**	**5.60**	0.928	6.31	0.913	6.72	0.918	6.60	0.908	6.73	0.926	5.88
Kanazawa	**0.807**	**13.31**	0.702	18.94	0.552	19.63	0.782	16.39	0.682	20.39	0.757	14.88
Hososhima	**0.910**	**4.81**	0.892	4.94	0.893	5.46	0.896	4.97	0.851	6.03	0.885	5.25

Table 4. *The indices of waves forecasted by the models for T_s.*

Station	**FS_XGBoost**		**CNN**		**LSTM**		**SVM**		**ANN**		**RF**	
	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)
Onahama	**0.742**	**11.83**	0.683	13.87	0.694	14.41	0.648	13.46	0.607	14.56	0.708	12.48
Kuzi	**0.703**	**13.60**	0.690	14.02	0.638	19.87	0.623	19.6	0.620	16.50	0.682	14.04
Akita	**0.841**	**9.40**	0.804	12.07	0.804	12.07	0.828	9.58	0.788	11.04	0.809	10.15
Kanazawa	**0.902**	**8.57**	0.762	13.46	0.683	27.38	0.845	11.18	0.704	21.71	0.882	9.61
Hososhima	**0.829**	**12.54**	0.803	13.26	0.812	13.22	0.827	13.15	0.710	17.17	0.779	13.37

Table 5. *Comparison of the XGBoost with feature selection to the original one on 5 stations.*

Station	H_p				T_s			
	FS_XGBoost		**XGBoost**		**FS_XGBoost**		**XGBoost**	
	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)	CC	NRMSE (%)
Onahama	**0.943**	**5.54**	0.928	6.10	**0.742**	**11.83**	0.733	12.44
Kuzi	**0.951**	**6.64**	0.950	6.73	**0.703**	**13.60**	0.673	14.15
Akita	**0.933**	**5.60**	0.928	5.80	**0.841**	**9.40**	0.833	9.63
Kanazawa	**0.807**	**13.31**	0.755	14.88	**0.902**	**8.57**	0.865	10.8
Hososhima	**0.910**	**4.81**	0.903	4.90	**0.829**	**12.54**	0.820	12.69

5. CONCLUSIONS

This study presents a nearshore wave forecasting framework using XGBoost with feature selection (FS_XGBoost), developed for five coastal stations in Japan. The models were trained on observed nearshore wave data and real-time global forecasts from JMA, NOAA, and ECMWF. For each station, separate XGBoost models were built for wave height and wave period prediction. Feature importance analysis identified key ECMWF inputs at various lead times (from real-time to 7 days ahead) as the most influential. FS_XGBoost outperformed other machine learning and deep learning models in predictive accuracy. The framework enables reliable weekly forecasts and is adaptable to any location with available nearshore wave observations.

Acknowledgement: *The authors would like to express our grateful thanks to the data support from Professor Sooyoul Kim works at the Center for Water Cycle Marine Environment and Disaster Management, Kumamoto University, Japan.*

REFERENCES

[1]. Z. &. D. A. Wei, *"A convolutional neural network based model to predict nearshore waves and hydrodynamics,"* Coastal Engineering, vol. 171, (2022).

[2]. P. P. J. C. M. D. R. & M. S. Bento, *"Ocean wave power forecasting using convolutional neural networks,"* IET Renewable Power Generation, vol. 15, no. 14, pp. 3341-3353, (2021).

[3]. S. H. J. L. Y. L. G. B. F. &. B. Z. Gao, *"A forecasting model for wave heights based on a long short-term memory neural network",* Acta Oceanologica Sinica, vol. 40, no. 1, pp. 62-69, (2021).

[4]. F. C. & F. L. Minuzzi, *"A deep learning approach to predict significant wave height using long short-term memory",* Ocean Modelling, vol. 181, (2023).

[5]. S. X. N. & D. S. Fan, *"A novel model to predict significant wave height based on long short-term memory network,"* Ocean Engineering, vol. 205, (2020).

[6]. S. A. S. G. N. G. F. &. S. J. J. Emmanouil, *"Statistical models for improving significant wave height predictions in offshore operations,"* Ocean Engineering, vol. 206, (2020).

[7]. D. S. D. P. S. R. H. S. &. S. A. Adytia, *"A Deep Learning Approach for Wave Forecasting Based on a Spatially Correlated Wind Feature, with a Case Study in the Java Sea, Indonesia",* Fluids, vol. 7, no. 1, (2022).

[8]. X. R. H. S. J. T. X. L. a. J. L. Q. Lin, *"A Hybrid Feature Selection Method Based on Imbalanced Learning for Wave Prediction,"* in IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA), Kaifeng, China, (2024).

[9]. S. T. T. H. T. M. &. M. H. Kim, *"A framework for transformation to nearshore wave from global wave data,"* Ocean Engineering, vol. 221, (2021).
[10]. N. K. S. R. & A. M. A. Kumar, *"Ocean wave height prediction using ensemble of extreme learning machine,"* Neurocomputing, vol. 277, pp. 12-20, (2018).
[11]. J. &. X. X. Chen, *"Ensemble learning based approach for the prediction of monthly significant wave heights,"* Renewable Energy, (2025).
[12]. F. Z. Y. C. B. &. J. S. C. O'Donncha, *"Ensemble model aggregation using a computationally lightweight machine-learning model to forecast ocean waves."* Journal of Marine System, vol. 199, (2019).
[13]. X. L. Y. G. S. &. R. P. Zhang, *"Ocean wave height series prediction with numerical long short-term memory,"* Journal of Marine Science and Engineering, vol. 9, no. 5, p. 514, (2021).
[14]. Makarynskyy, *"Improving wave predictions with artificial neural networks,"* Ocean Engineering, vol. 31, no. 5-6, pp. 709-724, (2004).
[15]. A. C. R. I. G. & R. J. R. Castro, *"Performance of artificial neural networks in nearshore wave power prediction,"* Applied Soft Computing, vol. 23, pp. 194-201, (2014).
[16]. J. &. M. E. A. Mahjoobi, *"Prediction of significant wave height using regressive support vector machines,"* Ocean Engineering, vol. 36, no. 5, pp. 339-347, (2009).
[17]. Z. &. D. A. Wei, *"A convolutional neural network based model to predict nearshore waves and hydrodynamics,"* Coastal Engineering, vol. 171, (2022).
[18]. J. X. B. &. W. J. Liu, *"Ensemble-based assimilation of wave model predictions: Contrasting the impact of assimilation in nearshore and offshore forecasting at different distances from assimilated data,"* Applied Ocean Research, vol. 140, (2023).
[19]. S. &. B. S. Adhikary, *"Improved Large-Scale Ocean Wave Dynamics Remote Monitoring Based on Big Data Analytics and Reanalyzed Remote Sensing,"* Nature Environment & Pollution Technology, vol. 22, no. 1, (2023).
[20]. A. &. Y. I. R. Takbash, *"Long-term and seasonal trends in global wave height extremes derived from era-5 reanalysis data,"* Journal of Marine Science and Engineering, vol. 8, no. 12, (2020).
[21]. F. C. &. F. L. Minuzzi, *"A deep learning approach to predict significant wave height using long short-term memory,"* Ocean Modelling, vol. 181, (2023).
[22]. J. L. F. Q. X. W. Y. S. J. S. C. &. Z. C. Zhang, *"Improving wave height prediction accuracy with deep learning,"* Ocean Modelling, vol. 188, (2024).

TÓM TẮT

Một mô hình học tổng hợp kết hợp với lựa chọn đặc trưng để dự báo sóng gần bờ

Nghiên cứu đề xuất một khung mô hình dự báo sóng gần bờ (OWFNW) tổng hợp trước một tuần để hỗ trợ viện quản lý vận chuyển và xây dựng tại các công trường hàng hải. Khung mô hình này sử dụng XGBoost với lựa chọn đặc trưng (FS_XGBoost) để dự báo tại 5 trạm trên bờ biển Nhật Bản. Các mô hình sóng dựa trên XGBoost được phát triển cho từng trạm, chuyển đổi dữ liệu sóng toàn cầu thành dự đoán sóng gần bờ. Các mô hình được huấn luyện bằng bốn bộ dữ liệu khác nhau từ Cơ quan Khí tượng Nhật Bản (JMA), Cơ quan Quản lý Khí quyển và Đại dương Quốc gia (NOAA), Trung tâm Dự báo Thời tiết Tầm trung Châu Âu (ECMWF) và Mạng lưới thông tin Sóng đại dương Toàn quốc dành cho Cảng và Bến cảng (NOWPHAS). Kết quả cho thấy việc lựa chọn các đặc trưng sẽ nâng cao độ chính xác dự đoán của mô hình và giảm thiểu các lỗi dự đoán. Phương pháp luận này cũng có thể được áp dụng cho các vùng biển gần bờ khác, nơi có dữ liệu dự báo sóng toàn cầu.

Từ khoá: Lựa chọn đặc trưng; XGBoost; Sóng gần bờ; Dự báo sóng toàn cầu.