AUTOMATIC FEATURE SELECTION FOR STORM SURGE FORECASTING

Hien Nguyen Thi[1,*], *Thi-Thu-Hong Phan*[2], *Sooyoul Kim*[3]

Abstract

Accurate prediction of storm surge events is crucial due to their severe nature, with the potential for widespread flooding and destruction. Timely and precise predictions are necessary to facilitate effective evacuation orders and enable prompt emergency responses. Machine learning has consistently demonstrated its superiority over traditional models in storm surge forecasting in terms of both accuracy and timeliness. However, the accuracy of machine learning models in storm surge prediction greatly relies on selecting the appropriate input features for training. To address this issue, this paper proposes a feature selection method that utilizes a GA to determine the optimal set of input features for machine learning models in storm surge prediction. Through experiments conducted on storm surge data from the Tottori coast of Japan and Vietnam, the proposed approach exhibits significant improvements in the accuracy of machine learning models for storm surge prediction.

Index terms

Feature selection; genetic algorithm; wrapper method; storm surge.

1. Introduction

A storm surge is a sudden rise in sea level along the shore, caused by a powerful storm such as a hurricane or a typhoon [1]. The surge is caused by a combination of factors, including the low pressure in the center of the storm, strong winds, wave-induced radiation stress, and the rotation of the Earth. As the storm approaches the shore, the winds push water toward the coast, and the low pressure causes the water level to rise even further. Storm surges can vary in height depending on the size and strength of the storm, as well as the shape of the coastline and the depth of the ocean floor. They can result in flooding of low-lying coastal areas, which can cause significant damage to buildings and infrastructure, and can also pose a serious threat to human life [1], [2].

Therefore, storm surge prediction in advance is crucial for effective disaster preparedness and response [1], [3]. Predicting storm surges before they occur allows authorities to issue timely and accurate evacuation orders, which can mitigate fatalities

[1]Institute of Information and Communication, Le Quy Don Technical University
[2]Department of Artificial Intelligence, FPT University, Da Nang
[3]Centre for Water Cycle, Marine Environment and Disaster Management, Kumamoto University
[*]Corresponding author, email: hiennt@lqdtu.edu.vn DOI: 10.56651/lqdtu.jst.v14.n01.1043.ict

and reduce the impact of the storm. Advanced storm surge prediction also provides time for communities to implement protective measures such as sandbags, levees, and floodgates, which can prevent or mitigate flooding and property damage. Furthermore, predicting storm surges in advance can help emergency responders and local officials allocate resources more effectively and strategically, ensuring that help and aid are delivered promptly to the areas most in need [1], [3].

Two main approaches to modeling storm surge prediction are process-based numerical models and machine learning (ML) models [1], [3]–[5]. Process-based numerical models rely on physical principles and mathematical equations to simulate the behavior of the ocean and atmosphere during a storm event. These models require a deep understanding of the underlying physics and are computationally intensive [1]. In contrast, ML models learn patterns and relationships from historical data and are less reliant on physical principles [6]. ML has emerged as a promising approach for storm surge prediction in recent years, and compared to process-based models, these models have many advantages [3]–[5]. First, they can identify complex patterns and relationships in the data that may be difficult for process-based models to capture. Second, they are computationally efficient, which can be advantageous for real-time storm surge forecasting and decision-making. Finally, ML models can be adapted and updated as new data becomes available, allowing for continuous improvement and refinement of the predictions [3], [7].

Feature selection is a crucial step in ML that involves identifying the most relevant and informative features or variables for a given task [8], [9]. In many cases, datasets contain a large number of features, and not all of these features may be relevant to the target variable or may introduce noise or redundancy. Including irrelevant or redundant features in the model can lead to overfitting, which results in poor generalization performance and reduced interpretability. Feature selection can help to identify the most informative features and remove the irrelevant or redundant ones, reducing the dimensionality of the data and improving the performance of the model. Moreover, feature selection can also help to reduce the computational cost of model training and evaluation, which is particularly important for large datasets or models with high-dimensional input spaces [8], [9].

Feature selection has been widely used to improve the accuracy of ML algorithms in hydro-meteorological forecasting problems [10], [11]. The feature selection approach has been shown to improve the accuracy of rainfall forecasting models, especially when dealing with large datasets with a high number of potential predictor variables [10], [12], [13]. By using a feature selection approach, temperature prediction models can also be improved in terms of accuracy and computational efficiency, as only the most important variables are included [14], [15]. Moreover, through the application of a feature selection approach, wind speed prediction models can be optimized to provide more accurate and efficient forecasts, benefiting a wide range of industries and activities, particularly in situations where wind conditions can have a significant impact on various activities such as aviation, transportation, and outdoor events [11], [16].

The selection of input features also greatly affects the accuracy of machine learning models for storm surge forecasting [4], [17]. In addition, for the same ML model for storm surge forecasting, each geographical location needs to select different input features. Moreover, for the same ML model used for storm surge forecasting at the same location but with different lead times, it is also necessary to select different sets of input features. For example, the study [4], [17] indicated that it is important to choose appropriate features to achieve the best performance in using artificial neural networks (ANN) for forecasting after-runner storm surges on the Tottori coast of Japan. For a 5-hour lead time surge level forecast, the best results are obtained when using input features such as surge level, sea-level pressure, a drop of sea-level pressure, longitude and latitude of the typhoon, sea surface level, and wind speed. However, for a 24-hour lead time surge level forecast, in addition to the input features needed for the 5-hour lead time surge level forecast, the wind direction input feature is also required [4], [17].

Thus, it can be seen that selecting a suitable set of input features for machine learning models to perform storm surge forecasting is crucial. Previous studies have also selected suitable features for machine learning models when performing storm surge forecasting for each location, but this selection is entirely based on experience and has not been automated [4], [17]–[20]. However, since each location will require different feature sets, relying on experience to select features for machine learning models will be time-consuming and not applicable to many locations. Hence, this paper proposes an automated method for selecting features for machine learning models to perform storm surge forecasting. The proposed method in the paper allows for the automatic selection of features for different machine learning models for various locations that require storm surge forecasting.

The rest of the paper is organized as follows. Section 2 reviews the background and the related work. Section 3 describes the proposed method. Section 4 summarizes the experimental setup. Section 5 presents and discusses the results. Finally, Section 6 concludes the paper.

2. Related works

2.1. Machine learning for storm surge forecasting

Machine learning techniques are being utilized to enhance the accuracy and efficiency of storm surge forecasting by developing models that learn from historical data, thereby facilitating timely disaster response [3]–[5], [21]. ANN are widely used for storm surge forecasting [4], [17], [21]–[26], consisting of interconnected nodes representing mathematical functions. Data preparation involves preprocessing historical storm surge data, which is then used to train the ANN model, adjusting weights and biases for accurate predictions [4], [17], [23], [26], [27]. Machine learning methods like decision trees (DT), support vector machines (SVM), and random forests (RF) are used for storm surge forecasting [7], [28]. Each method has its strengths and weaknesses, and the choice depends on the forecasting problem's requirements. These methods can improve accuracy and efficiency, leading to better disaster management and response.

Genetic algorithm (GA): GAs are optimization techniques that utilize natural selection and evolution to solve complex problems. They create a population of potential solutions and apply genetic operators to evolve them, effectively solving various fields [29], [30]. A GA involves initialization, evaluation, selection, crossover, mutation, and termination. It generates an initial population of potential solutions, evaluates each solution's fitness, selects the fittest individuals for the next generation, and applies genetic operators to generate new offspring. The crossover and mutation operators work together in an iterative process, ending when a satisfactory solution or maximum evaluations are reached [29], [30]. GAs are advantageous for their ability to efficiently search large solution spaces, handle complex and non-linear problems without explicit mathematical models, and can be parallelized and adapted to solve multi-objective optimization problems. They can generate solutions representing trade-offs between competing objectives, making them useful in real-world optimization problems where multiple objectives must be considered [29], [30].

2.2. Feature selection

Feature selection is a process to identify relevant subsets of features from a dataset, aiming to eliminate redundant or irrelevant ones, but it can be challenging due to the exponential increase in feature search space [8], [9], [31]. Feature selection involves a search method and an evaluation measure to identify relevant and informative features. The search method explores possible subsets, while the evaluation measure scores each subset based on its performance improvement. The goal is to maximize the evaluation measure for improved accuracy, efficiency, and generalization [8], [9]. Traditional and evolutionary search methods for feature selection are computationally efficient and easy to implement, but they may not identify complex relationships or produce suboptimal subsets. Evolutionary methods, like GAs and particle swarm optimization, can identify complex relationships and find optimal subsets for learning algorithms, but they can be computationally expensive and require parameter tuning [9]. Evaluation measures for features in feature selection are techniques used to assess the relevance or importance of variables in a dataset. They are typically classified into three categories: filter methods, wrapper methods, and embedded methods. Filter methods evaluate features based on statistical properties (e.g., variance, correlation, mutual information) independently of any model. They are fast and model-agnostic but may miss feature interactions. On the other hand, wrapper methods test feature subsets by training a specific model and measuring performance (e.g., accuracy). Wrapper methods have the potential to find the optimal feature subset for a specific learning algorithm, but they may be computationally expensive, particularly for large datasets with many features [31], [32]. Ultimately, embedded techniques incorporate feature selection inside model training, utilizing algorithms such as LASSO or feature significance derived from RF. They optimize speed and precision while utilizing

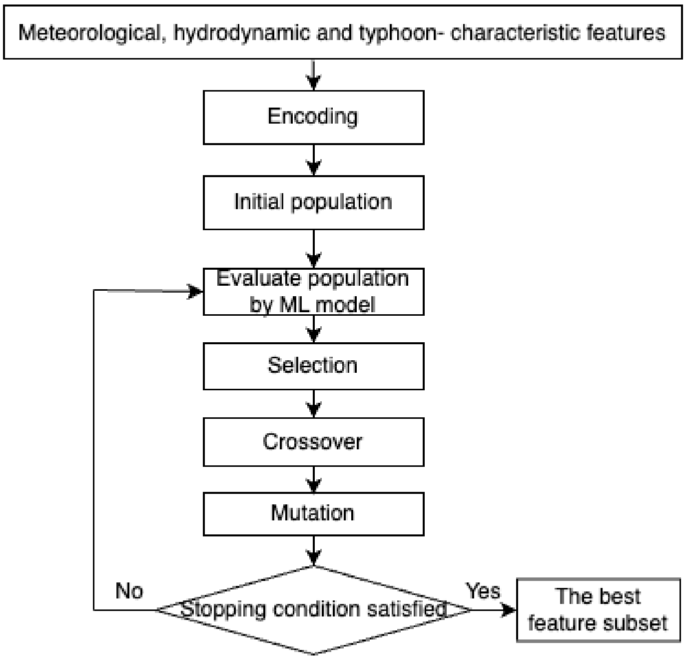


Fig. 1. The main steps of the proposed algorithm.

model-specific information. Each category possesses distinct advantages, rendering them appropriate for various situations in data preparation and model optimization.

3. Proposed method

The purpose of the proposed method is to find the optimal set of input features for a machine-learning model to predict storm surges. To achieve this goal, this paper proposes a GA-based wrapper feature selection approach. Figure 1 illustrates the main steps of the proposed method. The proposed method uses the GA to search for potential feature sets and a machine learning model to measure the quality of the feature set. The best feature set found by the proposed method is then used by the machine learning algorithm to build a prediction model for storm surges.

Algorithm 1 shows the pseudo code of the proposed method. The main components of the proposed method include chromosome representation, fitness evaluation, operators, and stopping conditions.

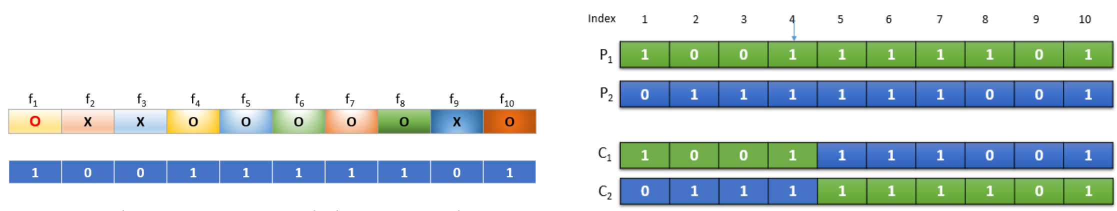
3.1. Chromosome representation

A chromosome is represented by a binary-encoded string with a length equal to the number of original features, where each cell represents a feature. Cells with a value of 1 indicate that the corresponding features are selected, while cells with a value of 0 indicate the corresponding features are not selected. Figure 2a shows an example of the

Algorithm 1: The pseudo-code of GA-based wrapper feature selection approach for storm surge forecasting

```
Input  : DT - training data with n original input features for storm surge prediction.
         DV - Validation data with n original input features for storm surge prediction.
         R - a regression learning algorithm to build a storm surge prediction model.
Output : The best feature subset.
Encode n features to chromosome and generate the initial population.
repeat
    /* Evaluate population by machine learning algorithm */
    for each chromosome in the population do
        sub ← decoding the chromosome to input features
        DT_sub ← DT reduced on sub features
        DV_sub ← DV reduced on sub features
        model ← R(DT_sub)
        Calculate RMSE on DV_sub
        Assign the RMSE as the fitness for the chromosome
    end
    Update the best chromosome of the population
    /* Create a new population */
    Selection.
    Crossover.
    Mutation.
until Stopping condition satisfied;
The best feature subset ← decoding the best chromosome to input features.
return The best feature subset
```

encoding of a chromosome with 10 original input features. The chromosome reflects 7 selected features and 3 remaining features that are not selected.



(a) The representation of chromosome for feature selection based on GA

(b) An illustration of point crossover

Fig. 2. Representation of an individual and an operator of a GA.

3.2. Fitness evaluation

The proposed method employs a wrapper approach to evaluate the quality of feature subsets. This method evaluates feature subsets based on training and evaluating machine learning models. The machine learning algorithm used to evaluate feature subsets is also used to build models for storm surge prediction.

To evaluate the quality of each chromosome, the following main steps need to be performed: first, calculate the selected feature set from the chromosome. Then, build the training and validation data sets from the original data set by keeping only the selected features. Next, the storm surge prediction model will be built on the training data set and validated on the validation data set. The normalized root mean square error (NRMSE) is used to evaluate the quality of the storm surge prediction model.

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

where n is the size of the training set, $y_{pre,i}$ represents the predicted surge level and $y_{obs,i}$ represents the actual surge level for the i_{th} data point (time index).

3.3. Operators

Selection: tournament selection determines which individuals are the most suitable for reproduction.

Crossover: the proposed method uses the standard single-point crossover. An example is shown in Figure 2b. Consider the following two chromosomes as an example: P_1, P_2. Firstly, a crossover point must be chosen at random from 1 to 10 (chromosome length) by generating a random number between 1 and 10. Let's say the crossover point is number four. The two new offspring after crossover are C_1, C_2.

Mutation: the mutation operator is applied to each chromosome bit, and the bit is randomly replaced by either 0 or 1.

3.4. Stopping condition

The processes of evolution are carried out for a maximum number of generations in this method. The best string found in the last generation provides a solution to the feature selection problem.

4. Experimental design

This section presents the feature selection experiments for storm surge forecasting in Japan and Vietnam to illustrate the proposed approach. We will briefly explain six other machine learning methods tried out in our experiments. They are the Multi-Layer Perceptron (MLP), k-Nearest Neighbors (k-NN), RF, Support Vector

Regression (SVR), Convolutional Neural Network (CNN) and Gradient Boosting Regressor (GBR). Whereas MLP is a form of ANN used in the paper [4], k-NN is chosen for its simplicity and ability to capture local patterns in data without assuming a specific distribution. RF 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. GBR is chosen for structured data due to its ability to handle non-linearity, reduce overfitting, and provide interpretable feature importance. These methods offer a range of approaches − non-parametric (k-NN), ensemble learning (RF and GBR), and margin-based (SVR) − to compare performance for feature selection.

4.1. Benchmark problem and parameter settings

The proposed method is evaluated on the data introduced in [4], [17]. The following meteorological and hydrodynamic parameters were used for three typhoons from the Japan observation stations:

- Meteorological parameters: wind speed (m/s), wind direction (degree), sea-level pressure (hPa) (SLP) and drop of SLP from average SLP (1013 hPa),
- Hydrodynamic parameters: sea surface level (m) (the atmospheric tidal level the surge level) and surge level (m) (= observed sea surface level atmospheric tidal level) in Sakai Minato and,
- Typhoon-characteristic parameters: longitude and latitude (degree), central atmospheric pressure (hPa) and highest wind speed near the typhoon center (m/s).

The parameters mentioned above were input parameters. The output is the surge levels with the lead times of 5, 12 and 24 hours (h) are forecasted. The dataset was collected during Typhoons Maemi 2003, Songda 2004 and Megi 2004 and normalized in the range of -1 to 1. Ten features are comprised in the dataset which are denoted in Table 1. In addition, we also used datasets collected from 15 storms in Vietnam (at Hon Dau and Hon Ngu stations) with similar features. We collected data from 12 storms at Hon Dau station, as detailed here [33]. Additionally, we gathered data from 3 storms at Hon Ngu station including Podul (2019), Bebinca (2018) and Son Tinh (2018), using a similar method.

Six common regression learning methods including SVM, MLP, KNN and RF are used in the experiments. Python is used to conduct the experiments, the zoofs library [34] for deploying GA and the Scikit-learn package are utilized to apply these regression techniques. The parameters of the algorithms are obtained based on grid search which tries all hyper-parameter combinations within the specified range, as given in Table 2.

4.2. Model performance

To evaluate the performance of the ML methods for the task of forecasting surge levels, we compare predicted values with the *true* values based on two metrics including

Table 1. The original input features of the benchmark dataset [4], [17]

No	Descriptor	Feature
1	Sea surface level (SSL)	F_1
2	Sea-level pressure (SLP)	F_2
3	Drop of sea-level pressure (DSLP)	F_3
4	Wind speed (WS)	F_4
5	Wind direction (WD)	F_5
6	Latitude (LT)	F_6
7	Longitude (LG)	F_7
8	Central atmospheric pressure of typhoon (CAP)	F_8
9	Highest wind speed near typhoon center (HWS)	F_9
10	Surge level (SS)	F_{10}

NRMSE (as shown in (1)) and the correlation coefficient (CC).

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

where n is the size of the training set, $y_{pre,i}$ represents the predicted surge level and $y_{obs,i}$ represents the actual surge level for the i_{th} data point (time index).

5. Results and discussion

The proposed method is evaluated for storm surge forecasting at different times: 5 h-ahead, 12 h-ahead, and 24 h-ahead in two cases 1) using the selected features proposed by our method and 2) using the whole feature as input.

5.1. A case of features selection for storm surge forecast 5 h-ahead

Table 3 shows the features selected by the proposed approach to forecast the surge level in 5 h-ahead. The best subsets of features obtained by applying GA-RF, GA-KNN, GA-MLP, and GA-SVM, are $\{F_1, F_2, F_3, F_6, F_7, F_8, F_9, F_{10}\}$, $\{F_1, F_2, F_4, F_6, F_7, F_8, F_{10}\}$, $\{F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}\}$, and $\{F_1, F_2, F_3, F_4, F_6, F_7, F_8, F_9, F_{10}\}$, respectively. To evaluate the performance of the aforementioned machine learning models, the best feature subset identified by the proposed method (GA-RF, GA-KNN, GA-MLP, and GA-SVM) and all the features are used on the test set.

Table 3, Figure 3 and Figure 4 reveal that the performance of the ML algorithms using the selected features by the proposed method is consistently better than when using all the features. Specifically, when using the selected features by the proposed method (i.e. 7.1%, 7.2%, 5.7%, and 8.9%)[1], the NRMSE values generated by the machine learning models are smaller than those generated by these methods using all the features (7.2%, 7.9%, 8.1%, and 11.6%, respectively). The results on the CC index are also consistent with this observation. The CC values produced by the proposed method

[1](noted, GA-RF, GA-KNN, GA-MLP, and GA-SVM)

Table 2. Optimal parameters of ML models for feature selection

Method	Parameter	Value
SVM	SVM Type	epsilon-SVM
	Cost	9.6974
	Gamma	6.8399
	Kennel type	RBF
	Epsilon	0.001
MLP	Hidden layers	5
	Learning rate	0.05
	Momentum	0.001
	Epochs	10000
KNN	K	7
	Distance function	Euclidean
RF	Max depth	0 (unlimited)
	# features	$int(log_2(\#predictors)+1)$
	# iterations	100
CNN	Conv1D(64, 3, activation='relu')	
	MaxPooling1D(pool size=3)	
	Activation('relu')	
	Dense(32,activation='relu')	
	Flatten()	
	Dense(1)	
	loss=mean_squared_error, optimizer=adam, epochs=50, batch size=8	
GRB	Number of Estimators	100
	max depth	4
	min samples split	3
	learning rate	0.1
	loss	squared_error
GA	Population size	100
	Genome length	10
	Population type	Bit strings
	Number of generations	100
	Crossover probability	0.6
	Mutation probability	0.1
	Reproduction rate	10%
	Selection type	Tournament (size=3)

Table 3. The error and CC of the proposed and benchmark methods for the lead time 5 h

No	Models	Feature	NRMSE	CC
1	RF	All	7.2%	0.963
	GA-RF	F_1, F_2, F_3, F_6, F_7, F_8, F_9, F_{10}	7.1%	0.965
2	kNN	All	7.9%	0.960
	GA-kNN	F_1, F_2, F_4, F_6, F_7, F_8, F_{10}	7.2%	0.968
3	MLP	All	8.1%	0.958
	GA-MLP	F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}	**5.7%**	**0.980**
4	SVM	All	11.6%	0.941
	SVM-MLP	F_1, F_2, F_3, F_4, F_6, F_7, F_8, F_9, F_{10}	8.9%	0.944
5	GBR	All	7.7%	0.968
	GA-GBR	$F_1, F_3, F_4, F_5, F_7, F_8, F_9, F_{10}$	7.0%	0.957
6	CNN	All	7.3%	0.970
	GA-CNN	$F_1, F_3, F_4, F_7, F_8, F_9, F_{10}$	6.9%	0.973

Table 4. Comparison of the best forecasting surge level 5 h-ahead between our proposal and Kim's solution [4], [17]

No	Selector + Model	NRMSE	CC
1	GA-MLP	5.7%	0.980
2	MLP-B3 [4], [17]	7.7%	0.974

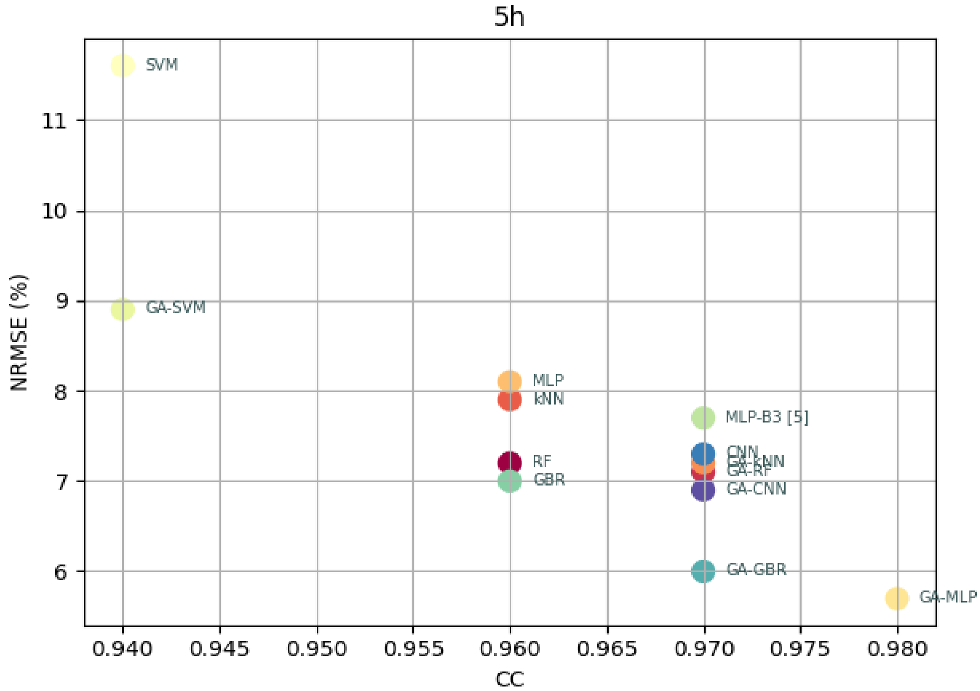


Fig. 3. Diagram of the CC and the NRMSE (%) for the surge level forecast made in 5 h in advance by 4 models and combination with GA-based features selection and the best features set in [4], [17].

Table 5. The error and CC of the proposed and benchmark methods for the lead time 12 h

No	Models	Feature	NRMSE	CC
1	RF	All	9.3%	0.961
	GA-RF	F_2, F_3, F_4, F_6, F_7, F_8, F_{10}	10.7%	0.947
2	kNN	All	10.3%	0.960
	GA-kNN	F_1, F_2, F_3, F_6, F_7, F_8, F_{10}	9.9%	0.965
3	MLP	All	8.2%	0.964
	GA-MLP	F_1, F_2, F_3, F_4, F_6, F_8, F_{10}	**7.6%**	**0.975**
4	SVM	All	7.9%	0.970
	GA-SVM	F_1, F_2, F_4, F_5, F_6, F_7, F_8, F_{10}	7.8%	0.970
5	GBR	All	9.7%	0.960
	GA-GBR	$F_1, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	8.1%	0.970
6	CNN	All	8.5%	0.970
	GA-CNN	$F_1, F_2, F_3, F_4, F_6, F_8, F_9, F_{10}$	7.7%	0.972

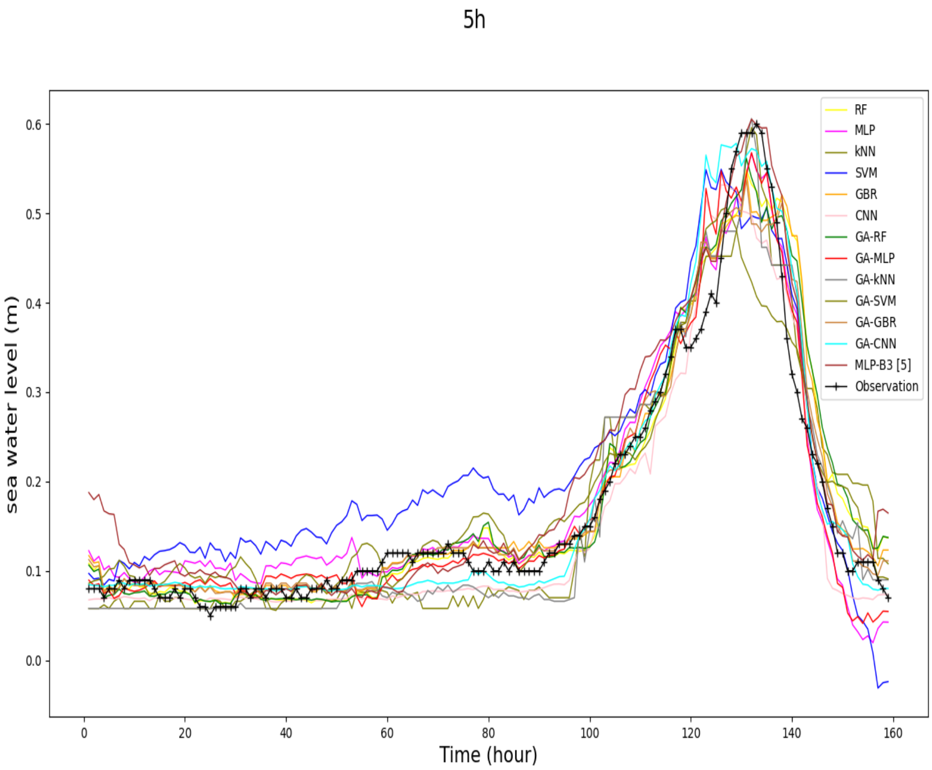


Fig. 4. Time series of observed and predicted surge levels for storms with the lead time of 5 *h at Sakai Minato by* 5 *models and wrapper frame.*

Table 6. Comparison the best GA-FS with NN feature selectors by Kim for lead time 12 *h*

No	Selector + Model	NRMSE	CC
1	GA-MLP	7.6%	0.975
2	MLP-B1 [4], [17]	11.2%	0.923

Table 7. The error and CC of the proposed and benchmark methods for the lead time 24 *h*

No	Models	Feature	NRMSE	CC
1	RF	All	10.1%	0.936
	GA-RF	F_3, F_6, F_{10}	13.1%	0.926
2	kNN	All	14.2%	0.900
	GA-kNN	F_1, F_3, F_4, F_6, F_8, F_{10}	13.5%	0.890
3	MLP	All	14.5%	0.910
	GA-MLP	F_2, F_4, F_5, F_6, F_8, F_9, F_{10}	**9.5%**	**0.964**
4	SVM	All	13.4%	0.893
	GA-SVM	F_1, F_2, F_3, F_5, F_6, F_7, F_8, F_9, F_{10}	11.0%	0.917
5	GBR	All	10.0%	0.929
	GA-GBR	$F_1, F_2, F_4, F_6, F_7, F_9, F_{10}$	9.9%	0.930
6	CNN	All	10.5%	0.921
	GA-CNN	$F_1, F_2, F_4, F_5, F_6, F_8, F_9, F_{10}$	10.0%	0.939

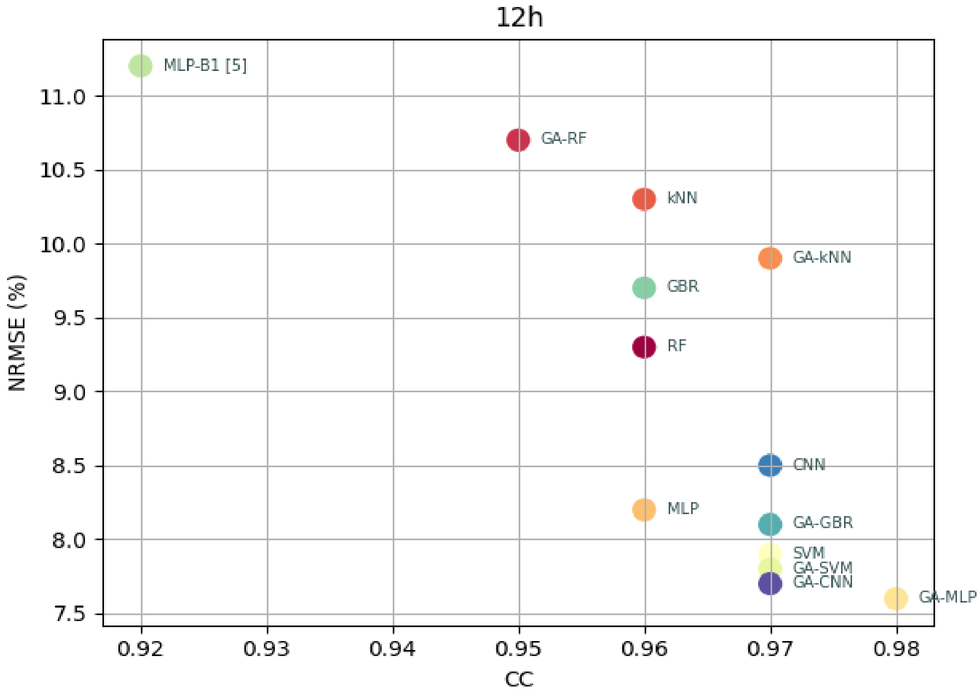


Fig. 5. Diagram of the CC and the NRMSE (%) for the surge level forecast made in 12 *h in advance by* 4 *models and combination with GA in wrapper features selection and the best features set in [4], [17].*

Table 8. Comparison the best GA-FS with NN feature selectors by Kim for lead time 24*h*

No	Selector + Model	NRMSE	CC
1	GA-MLP	9.5%	0.964
2	MLP-B4 [4], [17]	24.0%	0.940

(0.965, 0.968, 0.980, and 0.944) are superior to those generated by the corresponding baseline machine learning methods (0.963, 0.960, 0.958, and 0.941). These findings suggest that integrating the GA-based feature selection technique and machine learning methods yields better performance than using machine learning algorithms alone.

In addition to the quantitative assessment, this study also compares the visual prediction performance of the machine learning models. To assess the impact of the proposed method on the machine learning models, we visualize the prediction values of different methods (RF, GA-RF, KNN, GA-KNN, MLP, GA-MLP, SVM, and GA-SVM) with the actual observed values of surge level on the test set. As shown in Figure 4, the SVM method overestimates surge levels significantly when forecasting 5 h-ahead, while models GA-KNN slightly underestimate them. On the other hand, models RF, GA-RF, GA-KNN, and GA-MLP produce better predictive results, with predicted values fluctuating around the observed values. Some of these models, such as GA-KNN, have predicted surge levels that are close to the maximum surge level.

The MLP model using features extracted by GA-MLP produces the best prediction

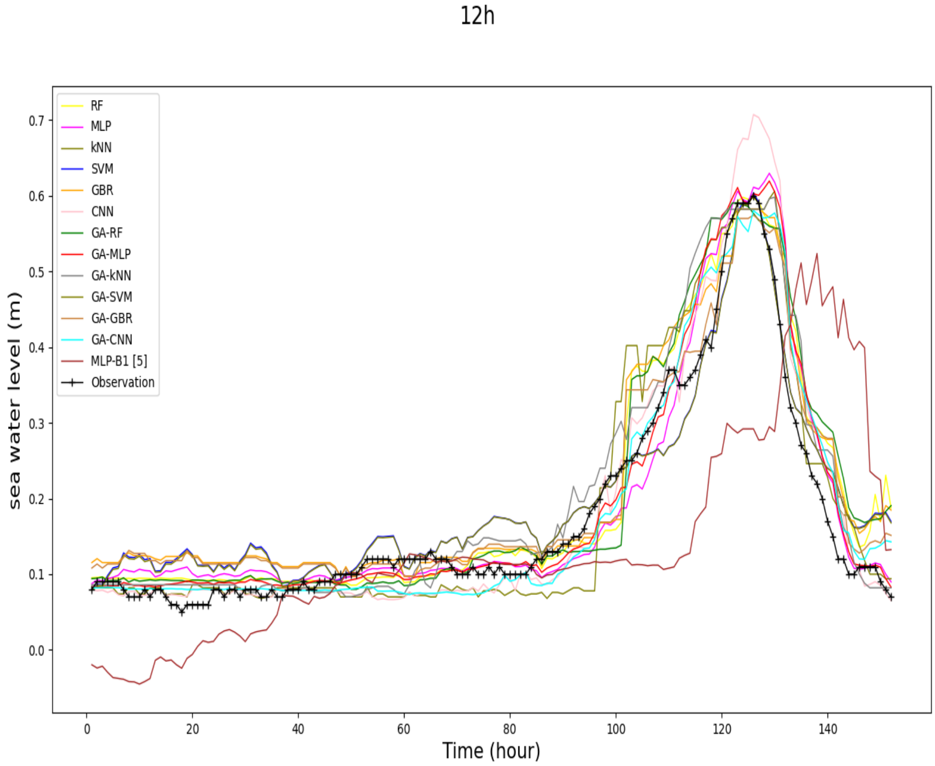


Fig. 6. Time series of observed and predicted surge levels for storms with the lead time of 12 *h at Sakai Minato by* 5 *models and wrapper frame.*

results among the machine learning methods, with a minimal NRMSE of 5.7% and a maximum CC of 0.980. To demonstrate the effectiveness of the proposed approach in forecasting surge level, we compare this result with the feature set B3 (including F_1, F_2, F_3, F_4, F_6, F_7, F_{10}) presented in the paper by Kim *et al.* [4], [17]. Table 4 compares the performance of the MLP model used in this study with that used in the study by Kim *et al.* [4], [17]. The feature set for MLP obtained automatically from the proposed method is different from the feature set B3 in [4], [17]. It clearly demonstrates a significant improvement in the MLP model, with its NRMSE decreasing from 7.7% to 5.5%, and its CC increasing from 0.974 to 0.980.

5.2. A case of features selection for storm surge forecast 12 h-ahead

Table 5 shows the features selected by the proposed approach for forecasting surge level in 12 h-ahead. The table reveals that the proposed approach outperforms the benchmark methods in terms of accuracy, except for RF. Among the machine learning methods, MLP utilizing features extracted by GA-MLP achieves the best prediction results, with a minimal NRMSE of 7.6% and a maximum CC of 0.975.

To assess the effectiveness of our proposed method, we compared our results with those of Kim's study in the best-case scenario: using the GA-based wrapper feature

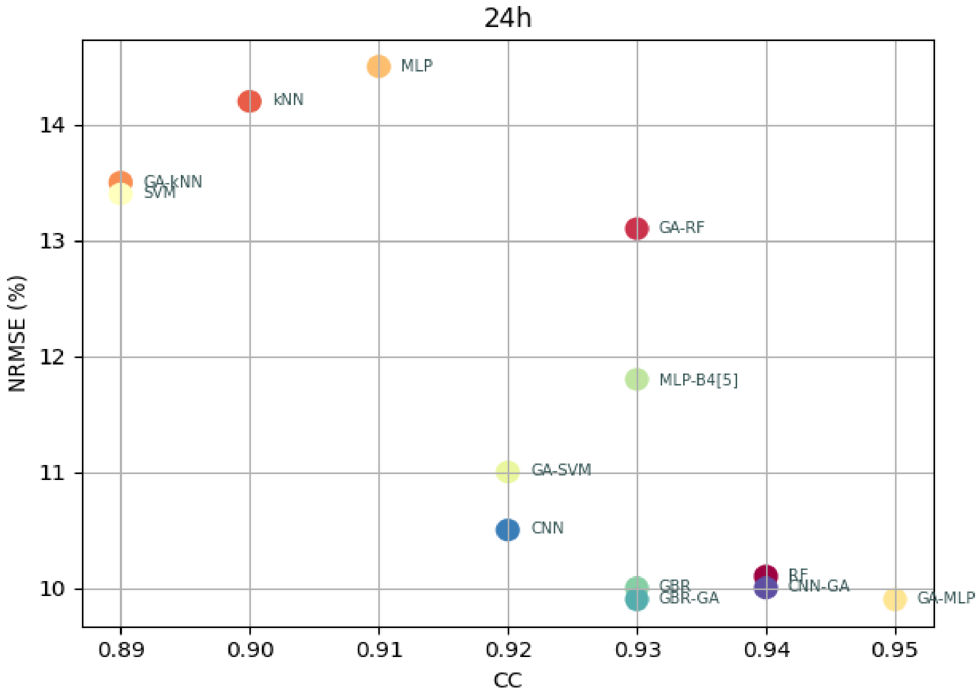


Fig. 7. Diagram of the CC and the NRMSE (%) for the surge level forecast made in 24 h in advance by 4 models and combination with GA in wrapper features selection.

extraction in combination with MLP, and with the best result obtained in [4], [17] (i.e. using feature set B1 which includes F_2, F_3, F_6, F_7, F_{10}). Table 6 clearly shows that the accuracy (measured by NRMSE) of the wrapper feature selection with GA using MLP is significantly better than that of the model-induced MLP using Kim's features. Moreover, as can be seen in Figure 6, the 12 h lead time forecasts of MLP-B1 (trained using feature set B1 in [4], [17]) greatly overestimate the surge levels, whereas GA-MLP predicts the surge levels with much less deviation.

From Table 3 and Table 5 and the representation of the values of these two tables in Figure 5 as well as time series of the surge forecasts and observations are displayed in Figure 6 it can be observed that the sets of selected features for the 12 h-ahead forecast are not the same as those for the 5 h-ahead forecast. Additionally, the machine learning methods combined with the GA-based feature extraction approach (GA-RF, GA-KNN, GA-MLP, and GA-SVM) do not perform as well when forecasting sea level 12 hours ahead as they do when forecasting 5 hours ahead.

5.3. A case of features selection for storm surge forecast with the lead time of 24 h

The selected features and accuracy of both the proposed and benchmark methods (using all features) are presented in Table 7. The table demonstrates that the model trained using the feature set automatically identified by the proposed approach performs better than the other baseline models, except for RF. This finding suggests that feature

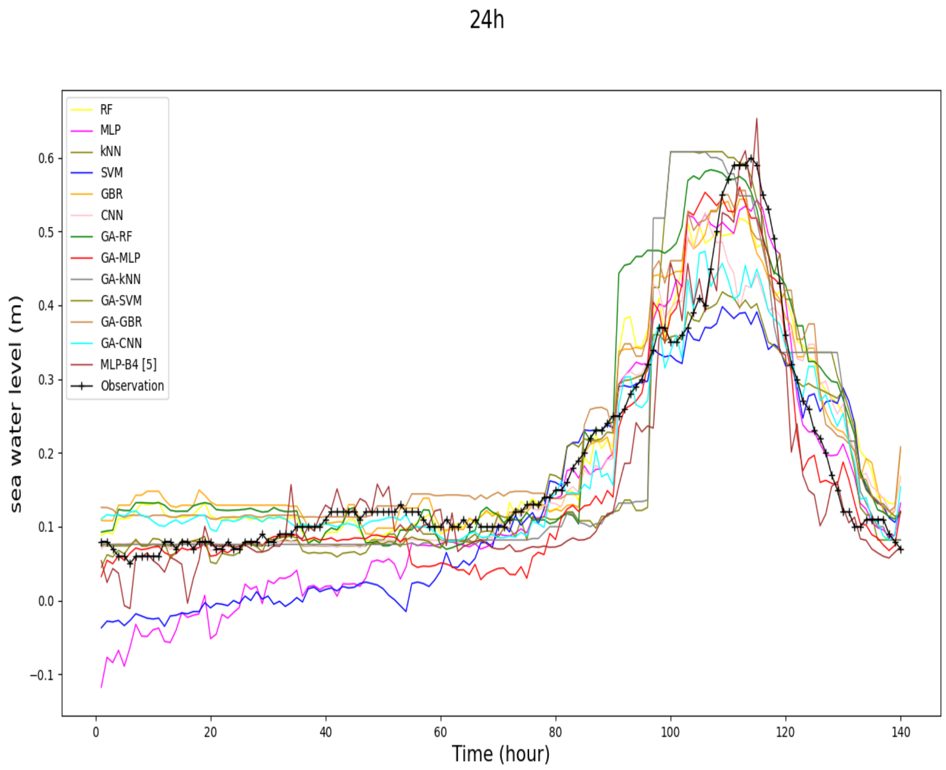


Fig. 8. Time series of observed and predicted surge levels for storms with the lead time of 24*h at Sakai Minato by* 5 *models and wrapper frame.*

selection using the GA-based method is more effective than relying solely on machine learning algorithms. Additionally, the results for the 24 h lead time demonstrate that the combination of wrapper feature extraction with GA and MLP produces the best prediction results.

Table 8 compares the results of the proposed wrapper feature extraction approach combined with MLP and the best feature set in [4], [17]. The results demonstrate that the accuracy (NRMSE) of the wrapper feature selection with GA using MLP is significantly better than the model-induced MLP using Kim's features. The feature set obtained automatically from the wrapper MLP and GA consists of 7 features, while the feature set B4 in [4], [17] has
8 features. Figure 7 shows the results in Table 7 and 8 while time series of the surge forecasts and observations for Sakai Minato are displayed in Figure 8.

5.4. A case of features selection for storm surge forecast in Vietnam

Data from 15 storms captured at the Hon Dau and Hon Ngu stations in Vietnam with 153 training samples and 76 testing samples were also used to further evaluate the methods. The dataset used 10 input features that were also used in the Japanese dataset. The results in Table 9 also show that using MLP with wrapper and GA techniques yields

Table 9. The error and the CC of the proposed and benchmark methods for the Vietnam data

No	Models	Feature	NRMSE	CC
1	RF	All	17.0%	0.755
	GA-RF	F_4, F_5, F_3, F_7, F_6, F_8, F_9, F_{10}	16.79%	0.766
2	kNN	All	19.7%	0.663
	GA-kNN	F_2, F_4, F_6, F_7	16.17%	0.808
3	MLP	All	18.1%	0.711
	GA-MLP	F_6, F_7, F_8, F_9, F_{10}, F_9, F_{10}	**14.9%**	**0.818**
4	SVM	All	18.2%	0.711
	GA-SVM	F_4, F_5, F_2, F_6, F_7, F_8, F_9, F_{10}	16.8%	0.794
5	GBR	All	16.2%	0.802
	GA-GBR	$F_5, F_6, F_7, F_8, F_{10}$	15.3%	0.809
6	CNN	All	18.0%	0.725
	GA-CNN	$F_1, F_2, F_4, F_5, F_6, F_8, F_9, F_{10}$	17.5%	0.743

the best results on this dataset; additionally, combining machine learning techniques with feature selection using GA-wrapper gives better results on this dataset than using only conventional machine learning techniques.

5.5. Comparison to other feature selection approaches

To compare the wrapper approach using GA with other wrapper techniques (Forward Selection and Backward Selection) and other approaches like filter approaches (including SelectKBest with f_regression, with mutual_info_regression and VarianceThreshold, Information Gain) or embedded techniques (including RF importance and Lasso regression), we conducted additional experiments (using MLP) with other feature selection techniques. The results are shown in Table 10. Evidently, the outcomes demonstrate that on all four datasets, MLP with GA-wrapper produces the best results.

6. Conclusion

The paper proposed the GA-based wrapper feature selection to find the optimal set of input features for a machine-learning model to predict storm surges. This method employs a GA to explore potential feature sets, and a machine-learning model is used to evaluate the quality of each feature set. The proposed method was compared to benchmark methods, including using all available features and manually selected features. The results showed that the proposed approach outperforms the benchmark methods in terms of prediction accuracy for all lead times (5 h, 12 h, and 24 h). In particular, the MLP model using features selected by the GA-based wrapper feature selection approach produced the best results in terms of minimal NRMSE and maximum CC. The results also demonstrated that the optimal feature set varied with lead time, and the GA-based wrapper feature selection approach was more effective

Table 10. Feature selection methods performance

	Filter				Wrapper		Embedded	
	KBest f_regression	KBest mutual_info	Variance Threshold	Information Gain	Forward Selection	Backward Selection	Random Forest Importance	Lasso
5 h								
NRMSE CC	9.82% 0.936	11.72% 0.894	17.76% 0.892	10.42% 0.934	12.31% 0.886	13.89% 0.845	11.68% 0.898	12.78% 0.922
Features selected	$F_1, F_2, F_3, F_4, F_5, F_6, F_8, F_{10}$	$F_1, F_2, F_3, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_1, F_2, F_3, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_2, F_3, F_4, F_5, F_7, F_9, F_{10}$	$F_1, F_2, F_3, F_4, F_5, F_7, F_9, F_{10}$	$F_1, F_3, F_4, F_5, F_6, F_7, F_8, F_{10}$	$F_5, F_6, F_7, F_2, F_3, F_4, F_{10}$	$F_3, F_4, F_5, F_6, F_8, F_9, F_{10}$
12 h								
NRMSE CC	16.12% 0.783	16.23% 0.846	18.87% 0.818	18.44% 0.872	18.09% 0.798	17.63% 0.745	15.55% 0.807	18.66% 0.719
Features selected	$F_1, F_2, F_3, F_4, F_5, F_6, F_8, F_{10}$	$F_1, F_2, F_6, F_3, F_5, F_7, F_8, F_9, F_{10}$	$F_1, F_2, F_3, F5, F_6, F_7, F_8, F_9, F_{10}$	$F_5, F_8, F_9, F_{10}, F_6$	$F_1, F_3, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_1, F_3, F_4, F_5, F_6, F_7, F_8, F_{10}$	F_1, F_2, F_3, F_6, F_7	$F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$
24 h								
NRMSE CC	19.03% 0.777	20.00% 0.729	20.90% 0.717	19.96% 0.704	21.12% 0.694	20.68% 0.721	22.07% 0.710	20.31% 0.714
Features selected	$F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_1, F_2, F_5, F_6, F_7, F_3, F_8, F_9, F_{10}$	$F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_2, F_3, F5, F_7, F_9, F_{10}$	$F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_2, F_3, F_5, F_6, F_7, F_9, F_{10}$	$F_3, F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$
Vietnam data								
NRMSE CC	18.4% 0.701	18.2% 0.711	17.9% 0.720	18.1% 0.712	17.8% 0.725	17.9% 0.720	18.0% 0.715	18.1% 0.710
Features selected	$F_1, F_2, F_3, F_4, F_8, F_9, F_{10}$	$F_1, F_2, F_3, F_4, F_5, F_6, F_{10}$	$F1, F_3, F_4, F_5, F_6, F_7, F_9, F_{10}$	$F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_4, F_5, F_6, F_7, F_8, F_9, F_{10}$	$F_2, F_4, F_5, F_6, F_7, F_9, F_{10}$	$F_1, F_2, F_3, F_6, F_8, F_9, F_{10}$	$F_1, F_2, F_3, F_6, F_7, F_8, F_9, F_{10}$

than using only machine learning algorithms or previously identified feature sets. Overall, the proposed approach offers a promising method for selecting the most informative features for predicting storm surges using machine learning models.

References

[1] N. Kohno, S. K. Dube, M. Entel, S. Fakhruddin, D. Greenslade, M.-D. Leroux, J. Rhome, and N. B. Thuy, "Recent progress in storm surge forecasting," *Tropical Cyclone Research and Review*, vol. 7, no. 2, pp. 128–139, 2018. DOI: 10.6057/2018TCRR02.04

[2] H. Von Storch and K. Woth, "Storm surges: perspectives and options," *Sustainability Science*, vol. 3, pp. 33–43, 2008. DOI: 10.1007/s11625-008-0044-2

[3] M. Tadesse, T. Wahl, and A. Cid, "Data-driven modeling of global storm surges," *Frontiers in Marine Science*, vol. 7, p. 260, 2020. DOI: 10.3389/fmars.2020.00260

[4] S. Kim, Y. Matsumi, S. Pan, and H. Mase, "A real-time forecast model using artificial neural network for after-runner storm surges on the Tottori coast, Japan," *Ocean Engineering*, vol. 122, pp. 44–53, 2016. DOI: 10.1016/j.oceaneng.2016.06.017

[5] J.-W. Lee, J. L. Irish, M. T. Bensi, and D. C. Marcy, "Rapid prediction of peak storm surge from tropical cyclone track time series using machine learning," *Coastal Engineering*, vol. 170, 2021. DOI: 10.1016/j.coastaleng.2021.104024

[6] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015. DOI: 10.1126/science.aaa84

[7] S. Rajasekaran, S. Gayathri, and T.-L. Lee, "Support vector regression methodology for storm surge predictions," *Ocean Engineering*, vol. 35, no. 16, pp. 1578–1587, 2008. DOI: 10.1016/j.oceaneng.2008.08.004

[8] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014. DOI: 10.1016/j.compeleceng.2013.11.024

[9] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2015. DOI: 10.1109/TEVC.2015.2504420

[10] A. Haidar and B. Verma, "A genetic algorithm based feature selection approach for rainfall forecasting in sugarcane areas," pp. 1–8, 2016. DOI: 10.1109/SSCI.2016.7849935

[11] X. Liu, H. Zhang, X. Kong, and K. Y. Lee, "Wind speed forecasting using deep neural network with feature selection," *Neurocomputing*, vol. 397, pp. 393–403, 2020. DOI: 10.1016/j.neucom.2019.08.108

[12] P. Das, D. Sachindra, and K. Chanda, "Machine learning-based rainfall forecasting with multiple non-linear feature selection algorithms," *Water Resources Management*, vol. 36, no. 15, pp. 6043–6071, 2022. DOI: 10.1007/s11269-022-03341-8

[13] A. Haidar and B. Verma, "A novel approach for optimizing climate features and network parameters in rainfall forecasting," *Soft Computing*, vol. 22, pp. 8119–8130, 2018. DOI: 10.1007/s00500-017-2756-7

[14] S. Duan, S. Huang, W. Bu, X. Ge, H. Chen, J. Liu, and J. Luo, "LightGBM low-temperature prediction model based on LassoCV feature selection," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–8, 2021. DOI: 10.1155/2021/1776805

[15] M. Jamei, M. Karbasi, O. A. Alawi, H. M. Kamar, K. M. Khedher, S. Abba, and Z. M. Yaseen, "Earth skin temperature long-term prediction using novel extended Kalman filter integrated with artificial intelligence models and information gain feature selection," *Sustainable Computing: Informatics and Systems*, vol. 35, 2022. DOI: 10.1016/j.suscom.2022.100721

[16] X. Kong, X. Liu, R. Shi, and K. Y. Lee, "Wind speed prediction using reduced support vector machines with feature selection," *Neurocomputing*, vol. 169, pp. 449–456, 2015. DOI: 10.1016/j.neucom.2014.09.090

[17] S. Kim, S. Pan, and H. Mase, "Artificial neural network-based storm surge forecast model: Practical application to Sakai Minato, Japan," *Applied Ocean Research*, vol. 91, 2019. DOI: 10.1016/j.apor.2019.101871

[18] T.-L. Lee, "Neural network prediction of a storm surge," *Ocean Engineering*, vol. 33, no. 3-4, pp. 483–494, 2006. DOI: 10.1016/j.oceaneng.2005.04.012

[19] T.-L. Lee, "Back-propagation neural network for the prediction of the short-term storm surge in Taichung harbor, Taiwan," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 1, pp. 63–72, 2008. DOI: 10.1016/j.engappai.2007.03.002

[20] M. Sztobryn, "Forecast of storm surge by means of artificial neural network," *Journal of Sea Research*, vol. 49, no. 4, pp. 317–322, 2003. DOI: 10.1016/S1385-1101(03)00024-8

[21] M. R. Hashemi, M. L. Spaulding, A. Shaw, H. Farhadi, and M. Lewis, "An efficient artificial intelligence model for prediction of tropical storm surge," *Natural Hazards*, vol. 82, pp. 471–491, 2016. DOI: 10.1007/s11069-016-2193-4

[22] E. Adeli, L. Sun, J. Wang, and A. A. Taflanidis, "An advanced spatio-temporal convolutional recurrent neural network for storm surge predictions," *Neural Computing and Applications*, vol. 35, pp. 18 971–18 987, 2022. DOI: 10.1007/s00521-023-08719-2

[23] S.-W. Kim, J. A. Melby, N. C. Nadal-Caraballo, and J. Ratcliff, "A time-dependent surrogate model for storm surge prediction based on an artificial neural network using high-fidelity synthetic hurricane modeling," *Natural Hazards*, vol. 76, pp. 565–585, 2015. DOI: 10.1007/s11069-014-1508-6

[24] D. Makarynska and O. Makarynskyy, "Predicting sea-level variations at the cocos (keeling) islands with artificial neural networks," *Computers & Geosciences*, vol. 34, no. 12, pp. 1910–1917, 2008. DOI: 10.1016/j.cageo.2007.12.004

[25] S. H. You and J.-W. Seo, "Storm surge prediction using an artificial neural network model and cluster analysis," *Natural hazards*, vol. 51, pp. 97–114, 2009. DOI: 10.1007/s11069-009-9396-x

[26] T. Wang, T. Liu, and Y. Lu, "A hybrid multi-step storm surge forecasting model using multiple feature selection, deep learning neural network and transfer learning," *Soft Computing*, vol. 27, no. 2, pp. 935–952, 2023. DOI: 10.1007/s00500-022-07508-8

[27] M. Bajo and G. Umgiesser, "Storm surge forecast through a combination of dynamic and neural network models," *Ocean Modelling*, vol. 33, no. 1-2, pp. 1–9, 2010. DOI: 10.1016/j.ocemod.2009.12.007

[28] N. T. Hien, C. T. Tran, X. H. Nguyen, S. Kim, V. D. Phai, N. B. Thuy, and N. Van Manh, "Genetic programming for storm surge forecasting," *Ocean Engineering*, vol. 215, 2020. DOI: 10.1016/j.oceaneng.2020.107812

[29] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, pp. 8091–8126, 2021. DOI: 10.1007/s11042-020-10139-6

[30] M. Seyedali, "Genetic algorithm," *Evolutionary Algorithms and Neural Networks: Theory and Applications*, pp. 43–55, 2019. DOI: 10.1007/978-3-319-93025-1_4

[31] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022. DOI: 10.1016/j.jksuci.2019.06.012

[32] L. Wang, S. Jiang, and S. Jiang, "A feature selection method via analysis of relevance, redundancy, and interaction," *Expert Systems with Applications*, vol. 183, pp. 1–8, 2021. DOI: 10.1016/j.eswa.2021.115365

[33] N. T. Hiền, T. T. Phúc, N. V. Mạnh, N. T. Quyên, and H. H. Vân, "Phương pháp dự báo nước biển dâng do bão dựa trên lập trình di truyền," *Vietnam Journal of Hydro-meteorology*, vol. 716, pp. 1–11, 2020. DOI: 10.36335/VNJHM.2020(716).2-11

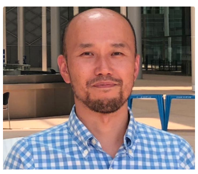
[34] Zoo Feature Selection - An open-source Library for feature selection. [Online]. Available: https://github.com/jaswinder9051998/zoofs

Manuscript received 31-12-2024; Accepted 15-6-2025.

■

Hien Nguyen Thi is a lecturer in the Department of Software Engineering, Institute of Information and Communication Technology, Le Quy Don Technical University. She received her Ph.D. degree from the Le Quy Don Technical University in 2015, her M.Sc. degree from the University of Engineering and Technology – Vietnam National University, Hanoi in 2007, and her B.Sc. degree from the Hanoi University of Science – Vietnam National University, Hanoi in 2000. Her research interests include machine learning, evolutionary computation, and search-based software engineering. Email: hiennt@lqdtu.edu.vn

Thi-Thu-Hong Phan received her Ph.D. in Automation, Computer Science, Image Processing, and Signal Processing from the University of Littoral, Calais, France, in October 2018. She obtained a Master's degree in Information Technology and Web (TIWe) from Claude Bernard Lyon 1 University, France, in June 2010, and a Bachelor's degree in Applied Mathematics and Informatics from the University of Natural Sciences, Hanoi National University, Vietnam, in June 2004. Her research interests include time series processing and forecasting, computer vision, machine learning/deep learning, and application of artificial intelligence to solving problems in agriculture. Email: hongptt11@fe.edu.vn



Sooyoul Kim is an associate professor in the Department of Civil and Environmental Engineering Education Program and the Disaster Mitigation Laboratory of the Centre for Water Cycle, Marine Environment and Disaster Management, Kumamoto University. He received his Ph.D. degree from Kyoto University in 2007, his Master of Engineering from Kyeonggi University in 2002. His research interests are data-based and physics-based modeling for predicting storm surges, waves, and tides. Email: sooyoulkim@kumamoto-u.ac.jp

TỰ ĐỘNG LỰA CHỌN ĐẶC TRƯNG CHO BÀI TOÁN DỰ BÁO NƯỚC BIỂN DÂNG SAU BÃO

Nguyễn Thị Hiền, *Phan Thị Thu Hồng*, *Sooyoul Kim*

Tóm tắt

Dự báo chính xác nước dâng do bão là rất quan trọng do tính chất nghiêm trọng của chúng, với khả năng gây ra lũ lụt và tàn phá trên diện rộng. Các dự báo kịp thời và chính xác là cần thiết để tạo điều kiện cho các lệnh sơ tán hiệu quả và cho phép phản ứng nhanh chóng khẩn cấp. Học máy đã liên tục chứng minh tính ưu việt của nó so với các mô hình truyền thống trong dự báo nước dâng do bão về cả độ chính xác và tính kịp thời. Tuy nhiên, độ chính xác của các mô hình học máy trong dự báo nước dâng do bão phụ thuộc rất nhiều vào việc lựa chọn các đặc trưng đầu vào phù hợp để huấn luyện. Để giải quyết vấn đề này, bài báo đề xuất một phương pháp lựa chọn đặc trưng sử dụng thuật toán di truyền để xác định tập hợp các đặc trưng đầu vào tối ưu cho các mô hình học máy trong dự báo nước dâng do bão. Thông qua các thí nghiệm được tiến hành trên dữ liệu nước dâng do bão từ bờ biển Tottori của Nhật Bản và Việt Nam, phương pháp đề xuất cho thấy những cải tiến đáng kể về độ chính xác của các mô hình học máy để dự báo nước dâng do bão.

Từ khóa

Lựa chọn đặc trưng; giải thuật di truyền; phương pháp wrapper; nước biển dâng do bão.