

## Significant wave height prediction through artificial intelligent mode decomposition for wave energy management

Yaoran Chen<sup>a</sup>, Dan Zhang<sup>b,c,\*</sup>, Xiaowei Li<sup>b,c</sup>, Yan Peng<sup>a</sup>, Chuhan Wu<sup>d</sup>, Huayan Pu<sup>b,c</sup>, Dai Zhou<sup>e</sup>, Yong Cao<sup>e</sup>, Jiujun Zhang<sup>f</sup>

<sup>a</sup> Collaborative Innovation Center for the Marine Artificial Intelligence, Shanghai University, Institute of Artificial Intelligence, Shanghai 200444, China

<sup>b</sup> School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China

<sup>c</sup> Engineering Research Center of Unmanned Intelligent Marine Equipment, Ministry of Education, Shanghai, 200444, China

<sup>d</sup> Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

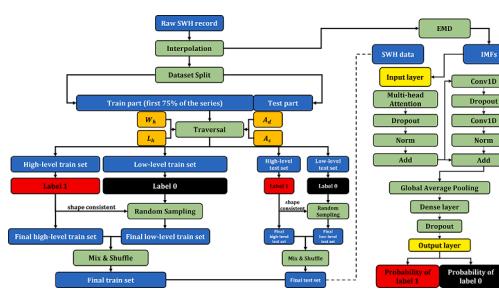
<sup>e</sup> School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

<sup>f</sup> College of Sciences, Institute for Sustainable Energy, Shanghai University, Shanghai 200444, China

### HIGHLIGHTS

- Significant wave height classification model is proposed based on Transformer.
- The empirical mode decomposition features can greatly enhance the model accuracy.
- The patterns of the best window size for the current model change with lead time.
- Parallel comparisons show the overall high-efficiency of the proposed model.

### GRAPHICAL ABSTRACT



### ARTICLE INFO

#### Keywords:

Significant wave height  
Time series classification  
Wave energy management  
Transformer  
Empirical mode decomposition

### ABSTRACT

The prediction of significant wave height (SWH) is crucial for managing wave energy. While many machine learning studies have focused on accurately predicting SWH values within hours in advance, the primary concern should be given to the level of the wave height for real-world applications. In this paper, a classification framework for the time-series of SWH based on Transformer encoder (TF) and empirical mode decomposition (EMD) is developed, which can provide a lead time of 6 to 48 h with the fixed thresholds of 2 m for high level waves and 1.5 m for low level waves. The performance of this approach is compared to that of three mainstream algorithms with and without EMD features. Results from the datasets collected from buoy measurements in the Atlantic Ocean indicate that the optimal mean accuracy at a lead time of 6 h was 99.1% and the average training time was 75 s, demonstrating the accuracy and efficiency of this proposed model. This study provides valuable tools and references for real-world SWH prediction applications.

E-mail address: [dan.zhang@shu.edu.cn](mailto:dan.zhang@shu.edu.cn) (D. Zhang).

\* Corresponding author at: Collaborative Innovation Center for the Marine Artificial Intelligence, Shanghai University, Institute of Artificial Intelligence, Shanghai 200444, China.

## Nomenclature

### Abbreviations and Symbols

SWH	Significant Wave Height
SWAN	Simulating Waves Nearshore
TCN	Temporal Convolutional Network
CNN	Convolutional Neural Network
RNN	Recurrent neural network
LSTM	Long Short-term Memory Network
GRU	Gated Recurrent Unit
STL	Seasonal Decomposition through Loess
EMD	Empirical Mode Decomposition
TSC	Time Series Classification
CDIP	Coastal Data Information Program
DPS	Decomposing-Predicting-Summing up routine

CCE	Categorical cross entropy
CA	Categorical accuracy
IoT	Internet of things
$H_s$	Significant wave height
$W_h$	Window size
$L_h$	Lead time
$A_d$	High-level SWH threshold
$A_s$	Low-level SWH threshold
$h_s$	Head size in multi-head attention
$h_n$	Head number in multi-head attention
$t_c$	Observation value under class $c$
$p_c$	Prediction value under class $c$
$C$	Number of classification types
$M_{test}$	Number of correct predictions in test set
$N_{test}$	Total number of samples in test set

## 1. Introduction

Ocean waves are one type of the significant sources of renewable energy, which can be converted into electricity through attenuators and other devices. They have the power flow intensities below the surface, which are in the range from 0.1 to 0.3 kW/m<sup>2</sup> to 2–3 kW/m<sup>2</sup> and one order of magnitude higher than wind energy on the surface. Additionally, ocean wave energy is more reliable than wind energy [1]. Although wave-power technology is not yet fully developed, it has a great potential for the future energy applications.

Regarding the ocean waves, their significant wave heights (SWHs) are the vital characteristics of the ocean environment [2]. Therefore, grasping and then predicting the advanced knowledge of SWHs are essential for effectively managing wave energy. In particular, when using machines to make energy conversion, the ability should be established to halt machines in the event of insufficient wave energy source or beyond the machines' capabilities [1]. To make efficient ocean energy conversion and utilization, theoretical modeling simulation is always needed. Traditional methods based on numerical simulations, such as the third-generation wave models (e.g., SWAN), require a high amount of computational power to solve Navier-Stokes equations using large oceanographic data sets. As a result, they are not suitable for short-term demands during the wave development processes, such as those with lead times of only a few hours [3].

Recently, statistical and machine learning approaches have provided potential solutions for the above challenge [4], and the core idea is to establish nonlinear patterns between consecutive time steps based on historical time series. So far, various time series models, including temporal convolutional network (TCN) [5], 1D convolutional neural network (CNN) [6] and the most prevalent models, have been developed based on recurrent neural networks (RNN) [7]. Generally, the models using convolutional networks display prominent superiority in time budget while models comprising RNN, such as long short-term memory network (LSTM) or gated recurrent unit (GRU), are more versatile, and present higher accuracy in various datasets [8]. Similar studies can also be referred in related fields, including wind speed forecasts [9], solar energy estimates [10], air-pollution monitoring [11] and Covid-19 spreading predictions [12].

In general, the specific height of the SWH with 2.2 m or 2.4 m may not be crucial in decision-making for wave energy management. The key factor is the potential for high energy harvest from "high wave level". Vice versa, the forecasts of 0.7 m or 0.9 m SWH values would generally result in the same strategies because both indicate that the sea state is "low". Then, by incorporating machine learning as the methodology, the regression problem can be transformed into a classification problem, while still utilizing historical time series as the input. Such a task is more beneficial to providing more precise and long-period predictions, while

regression models appeal to be inferior in those aspects because the efforts for forecasting the accurate values may not be necessary. However, at present, the related literatures are very limited.

For the time series classification (TSC) of SWH, besides the accuracy, at least other two important issues are included: the first one is the maximum lead time of the model with acceptable precision, and these two items are generally contradictory, that is, the capacity of the model is to hold its accuracy as the lead time is increasing, and the other is the training time cost, which determines whether the model could be dynamically updated in real world. Meanwhile, for tasks related to SWH where the records of measurements are difficult to be uploaded, especially in high sea states, the low training cost is also conducive to the edge computing through IoT devices or local edge servers.

Among prevalent models, including various CNN and RNN based models [5], the above two aspects of accuracy and time-cost cannot be balanced by purely adopting one of them. For instance, the CNN based models are hard to build long-term connections, and RNN based models need excessive budget to train. Hereby, the Transformer (TF) could be a promising alternative because of its attention mechanism and high-parallel capacity.

Also, according to previous studies, the series decomposition methods may help to solve above concerns. In the work of Yang et al. [6], the seasonal and remainder components were generated through Loess method (STL) that has outperformance in shallow water region, especially when lead time is longer than 24 h. In this regard, one study also proved that the STL required extremely limited decomposing time within 0.08 s. In literature [13], the wavelet transformation method was adopted to improve the ResNet, thus reducing up to 18% of mean squared error, and saving the training cost at the same time. The most popular decomposition is a totally posteriori method: empirical mode decomposition (EMD) where several studies connected EMD with different learning methods to largely improve the predicting accuracy [8]. Despite the positive effects as our past work suggested (such as the posteriori [14] and the self-adaption [15]), their common routine that the first is to decompose modes, then separately predict each and finally sum up all (abbreviated as DPS in this paper) is too intricate for large-dataset and real-time update [16]. Most importantly, such a DPS routine is indirect for the current SWH classification task, which hinders us to adopt a more suitable framework.

Therefore, in this study, a novel TSC framework is proposed for SWH prediction based on the transformer encoder (TF) [17], and the powerful signal decomposition method such as the empirical mode decomposition (EMD) [18]. Few recent studies from Pokhrel et al. [19] and Putri and Adytia [20] have preliminarily exhibited the ability of pure TF model on SWH predictions for solving differential equations [19] or handling a long lead time that contradicted to short-term scenario [20]. In this work, inheriting from our previous paper [16], the sub-series from EMD

are directly used as the features to further enhance the TF model for within 6 to 48 h in advance. Such framework can provide a more concise and effective way for building models that are specifically suited for time-series classification tasks. The discussion of the performance is elaborated through parallel model comparison in terms of the accuracy, robustness as well as the time-efficiency. Detailed information can be found in [Sections 3 and 4](#).

The structure of this paper is as follows. In [Section 2](#), the data source and data processing techniques are introduced; in [Section 3](#), the methods of the current TSC framework are presented; in [Section 4](#), the results are elaborated with multi-aspects discussion; and in [Section 5](#), solid conclusions together with the future works are given.

## 2. Dataset

The source of SWH observations and the pre-processing techniques are presented in this section. In [Section 2.1](#), the location and statistical exploration of the dataset are presented. In [Section 2.2](#), the key parameters of the task and the approaches for train and test data generation are illustrated.

### 2.1. Data source

The raw historical records of SWH are obtained from the widely used Coastal Data Information Program (<https://cdip.ucsd.edu/>) [21], which began its inception since 1975 under the funding of U.S. Army Corps of Engineers. As shown in [Fig. 1](#), the current offshore buoy is located at the Masonboro inlet of North Atlantic Ocean (34.14° N, 77.71° W) where the water depth is 15.700 m.

As shown in [Table 1](#), the dataset is a 4-year extraction with the time resolution of 30 min from year 2017 to 2020. It can be seen that the total number of timestamps is over 70,000, during which the global maximum SWH ( $H_s$ ) is 5.100 m that reaches a sea state of level-6, and the minimum value is 0.260 m. Both mean and the medium values are around 0.900 m, with a moderate standard deviation of 0.414 m.

Fortunately, the raw measurements are mostly completed, with only 4 missing. To complete the dataset, linear interpolation is used, and the resulting curve is shown in [Fig. 2](#).

**Table 1**

The statistical information of raw SWH buoy data from CDIP [21].

Number of points	Time step (min)	Mean (m)	Median (m/s)	Max (m)	Min (m)	Standard Deviation (m)
70,124	30	0.945	0.850	5.100	0.260	0.414

### 2.2. Data preprocessing

The TSC model requires careful data preparation. Therefore, in this section, the important definitions for current problem and the way for data generation are introduced as follows.

#### 2.2.1. Key parameters

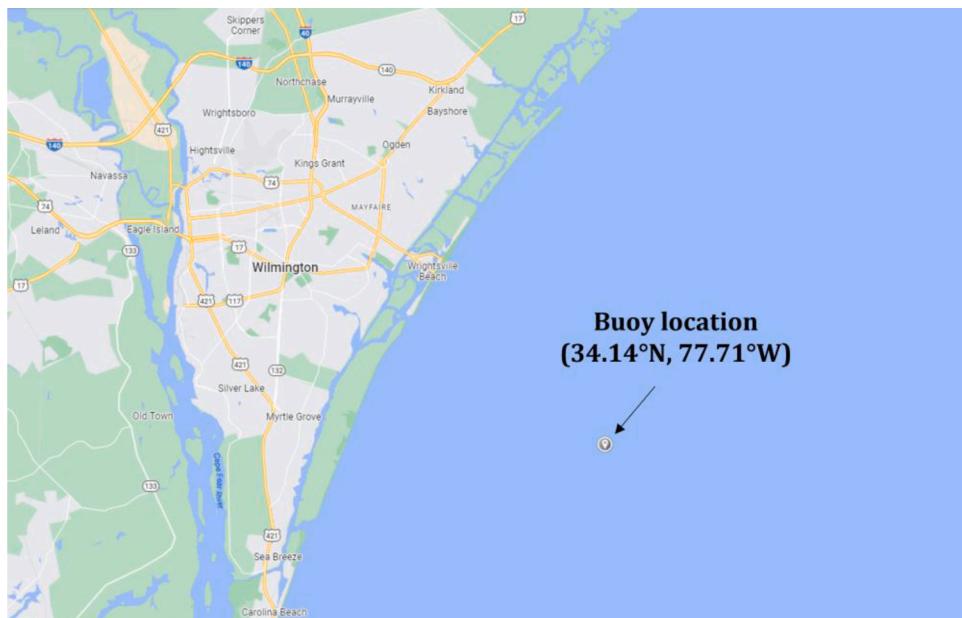
To begin with, four critical parameters of the task are introduced in this part, including window size  $W_h$ , lead time  $L_h$ , high-level SWH threshold  $A_d$  and low-level SWH threshold  $A_s$ . The graphical illustration is plotted in [Fig. 3](#).

As shown in [Fig. 3](#), the length of the solid green line is  $W_h$ , and all recorded points along the curve are known. The length of the dashed blue line is  $L_h$ , and the records are both unknown and unpredicted. The type (instead of the value) of the end point is as follows: if it is predicted to be larger or equal than  $A_d$  (horizontal dashed orange line), it is assumed to be “high-level” and will be painted with red ([Fig. 3\(a\)\)](#); otherwise, if it is predicted to be lower than  $A_s$  (horizontal dashed black line), it is assumed to be “low-level” and will be painted with black ([Fig. 3\(b\)\)](#). In the demo figure above, the  $W_h$  and  $L_h$  equal to 30 and 12 h, respectively, and their influences will be systematically investigated. The  $A_d = 2\text{m}$  and  $A_s = 1.5\text{m}$  are fixed in this study, which roughly represent the thresholds of medium/big wave and small wave according to Ref. [22].

#### 2.2.2. Dataset generation

Based on the interpolated raw measurements and the parameters above, the way of forming training and testing datasets is introduced in this part. The whole procedure is shown in [Fig. 4](#)

The complete series is split by time sequence: the first 75% is the train part and the rest 25% is the test part. Then, the two parts are separately traversed according to 4 key parameters that generate four sets, namely high-level train, high-level test, low-level train, and low-



**Fig. 1.** Buoy location for SWH measurements in Masonboro inlet near Wilmington [21].

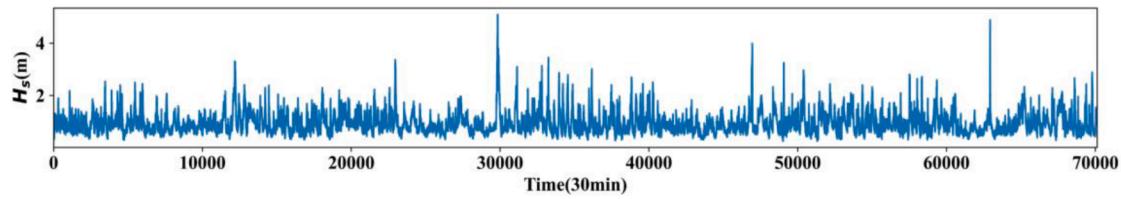


Fig. 2. The complete SWH dataset from CDIP (2017–2020) [21].

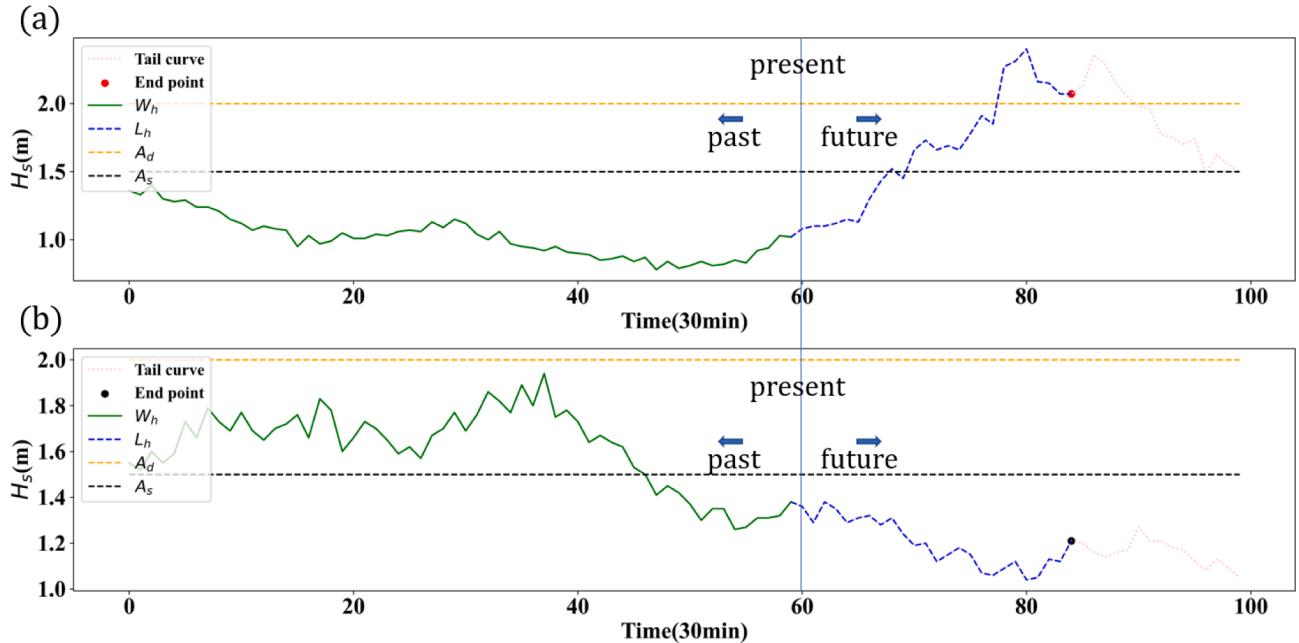


Fig. 3. The illustration of 4 key parameters for SWH classification: the green solid lines are known information and the red/black color of the end point represents the type of the wave level. (a) High-level ending; and (b) low-level ending.

level test. The first two (high-level) are labeled as 1 and the other (low-level) is labeled as 0. Next, to ensure balanced training, the number of high-level and low-level instances in the sample are made equal by down-sampling the low-level sets. The resulting high-level and low-level sets are then combined and shuffled before being used for machine learning.

### 3. Methodology

This work utilizes a deep learning model with a feature generation component using EMD and TF encoders, followed by a dense layer and the classification output. The overall framework is illustrated in Fig. 5.

The model structure is concise. The SWH series is first decomposed through EMD approach, thus resulting in several intrinsic mode functions (IMFs). Subsequently, the SWH and those sub-series are concatenated and undergo the pre-processing procedure, as explained in Section 2.2.2. The obtained training sets are fed into two TF encoders in succession for learning nonlinear patterns between time steps and features. The global average pooling is then used to reduce the tensor dimension into two for dense layer output. Eventually, two fully connected dense layers (one hidden and one output) are added to provide the probability matrix of the level of the wave, and an additional dropout layer is inserted for stronger generalization.

#### 3.1. Empirical mode decomposition

The EMD method was initially proposed by Huang et al. [18] in 1998, and has been widely implemented throughout decades for its

benefits in adaptively handling any nonlinear and nonstable signals [23]. The sub-components from EMD are called intrinsic mode functions (IMFs), with the following characters: (1) the difference between the number of local extrema and zero points should be zero or one; (2) the mean value of envelope curves of local maximum and local minimum should all be zero at any timestamp. As shown in Fig. 6, the total number of IMFs in this work is 15, and the last series is the residue term. The addition of them forms back to the original series.

As shown in Fig. 6 and compared with the original data series in Fig. 2, the IMFs appear to be more regular, partially because the mentioned characters can force to eliminate the long-term tendency fluctuations in each component, so the IMF becomes more stable. Also, during the decomposition process, as the frequency is gradually decreased from high to low, the local subtle stochastic terms that highly influence the prediction accuracy [16] are successfully extracted from the original data, which would help to greatly improve the TSC performance as well, which will later be shown in Section 4.

#### 3.2. TF encoder

In this subsection, the structure and process of the deep learning model, TF encoder, are briefly discussed. The TF was first proposed by Vaswani et al. [17] to solve natural language processing (NLP) problem with the attention mechanism tactfully involved initially. Recently, there have also been wave height related applications using TF models [20]. Generally, it comprises embedding layers to expanding the representations of word vectors, and both encoder and decoder parts were used for sequence-to-sequence mappings (such as the translation tasks).

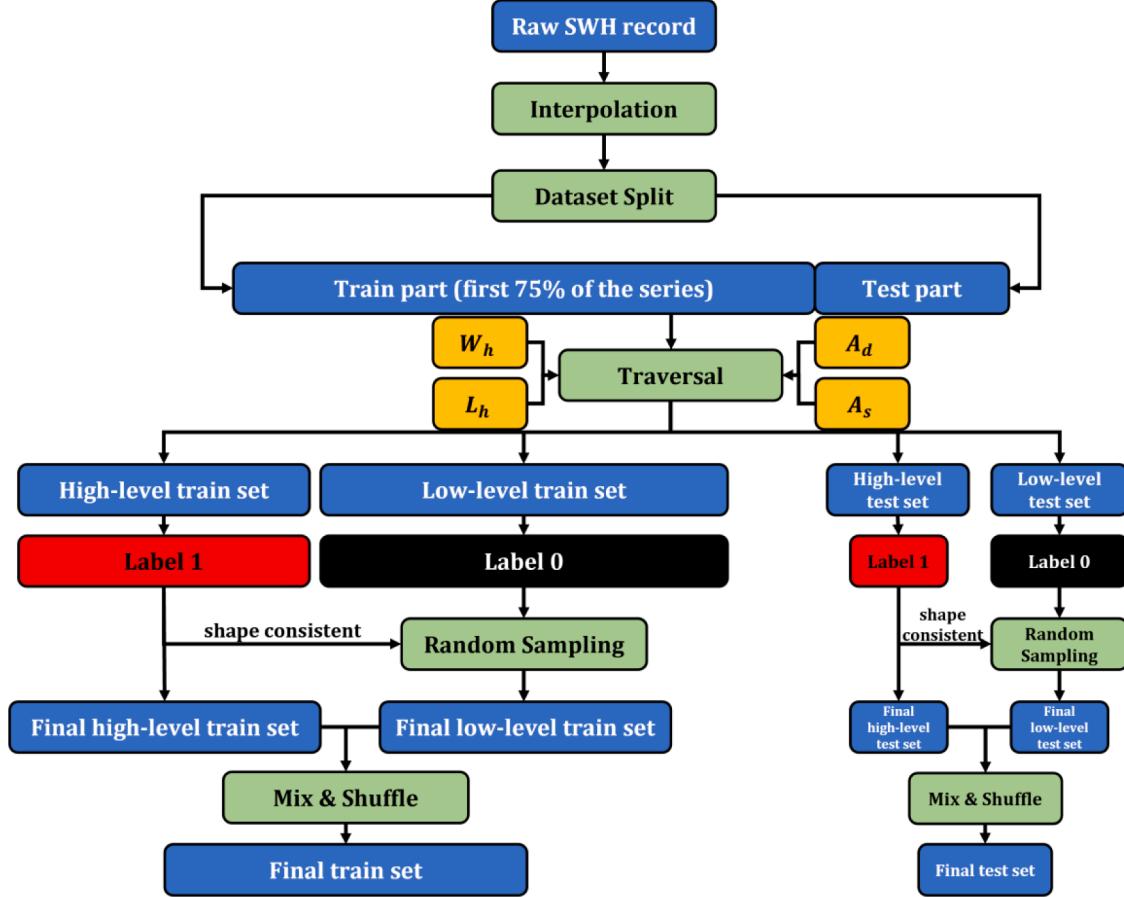


Fig. 4. The flowchart of dataset preparation for machine learning models.

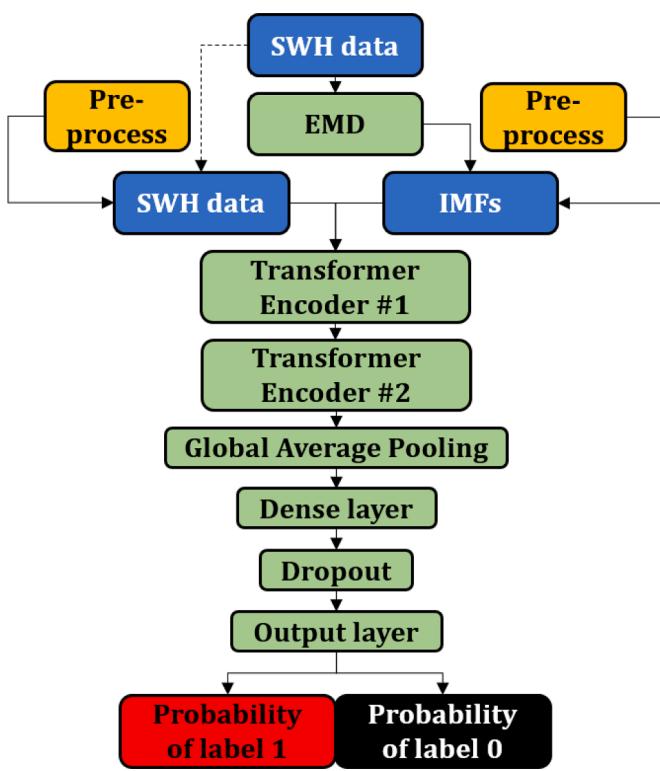


Fig. 5. The flowchart of machine learning model for SWH prediction.

However, in our work, as a binary classification task, only the encoder part is used.

As shown in Fig. 7, the left part is the structure of the self-attention module while the right part is the feed forward part. The multi-head attention layer calculates the attention score of inputs according to the following equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where  $Q$ ,  $K$ ,  $V$  are the matrixes of queries, keys and values, the initial values of which are identical; and  $d_k$  is the dimensional of the key matrix. As shown in Formula (1), the multi-head attention layer creates several independent linear representations from the  $Q$ ,  $K$ ,  $V$  inputs. These representations are then normalized and combined with the inputs before being passed to the feed-forward module on the right. In this study, the 1-D CNN is utilized to improve the representation of feature correlations.

### 3.3. Training settings

The key hyper-parameters of the model are presented and explained in this sub-section. Moreover, the training techniques and server setting are elaborated.

As shown in Table 2, both the head size  $h_s$  and the number of heads  $h_n$  are set to be 6 and 2. TF blocks in tandem are used to establish deeper connections between neurons. The first filter dimension for 1D-CNN is set to be multiplication of  $h_s$  and  $h_n$ , where the kernel size is quite small and set to be one window per kernel.

The Adam [24] was adopted as the optimizer, with an initial learning of  $10^{-3}$ . The activation function except the output layer is set to be the

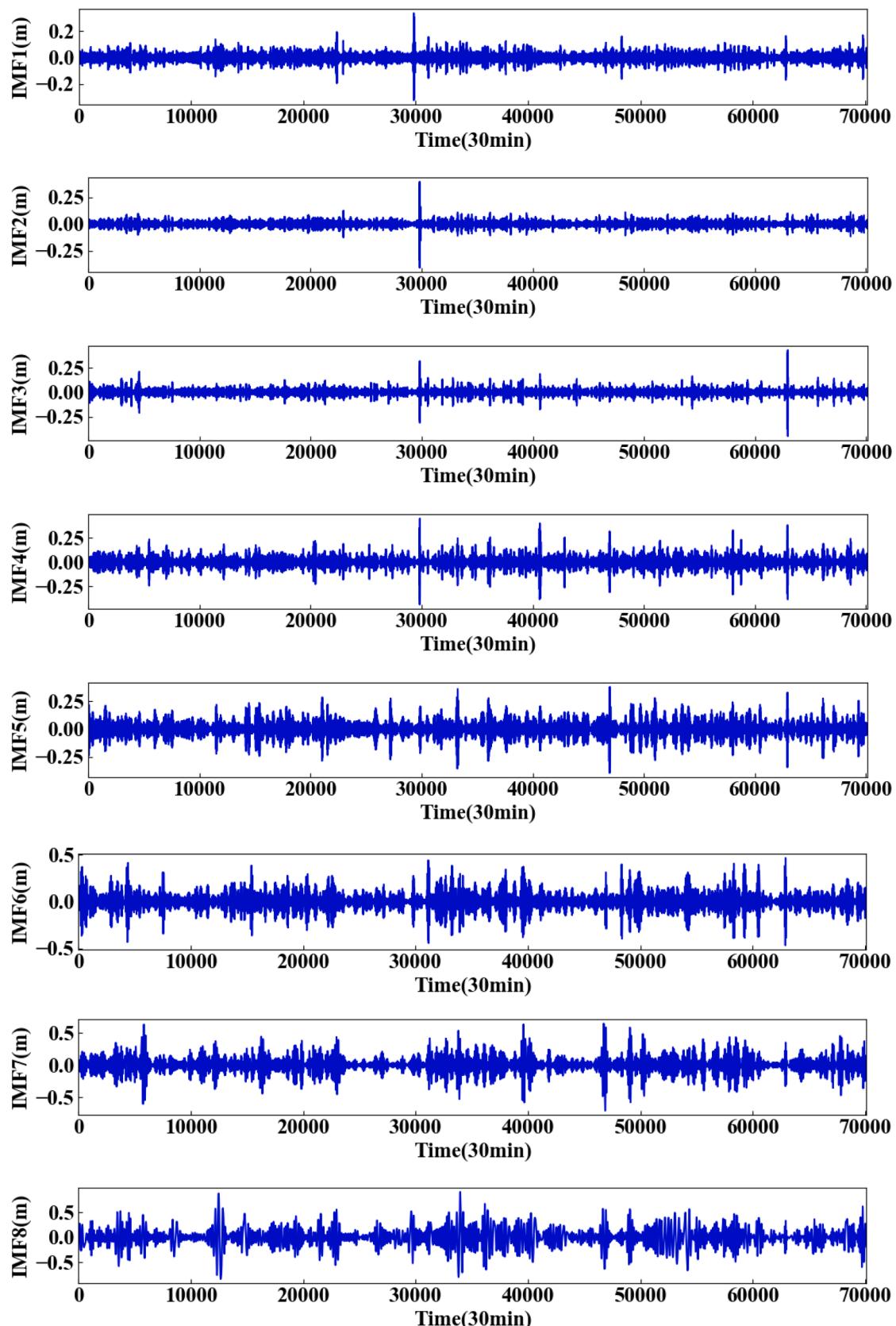


Fig. 6. All 15 intrinsic mode functions and the residue of the SWH series after EMD processing.

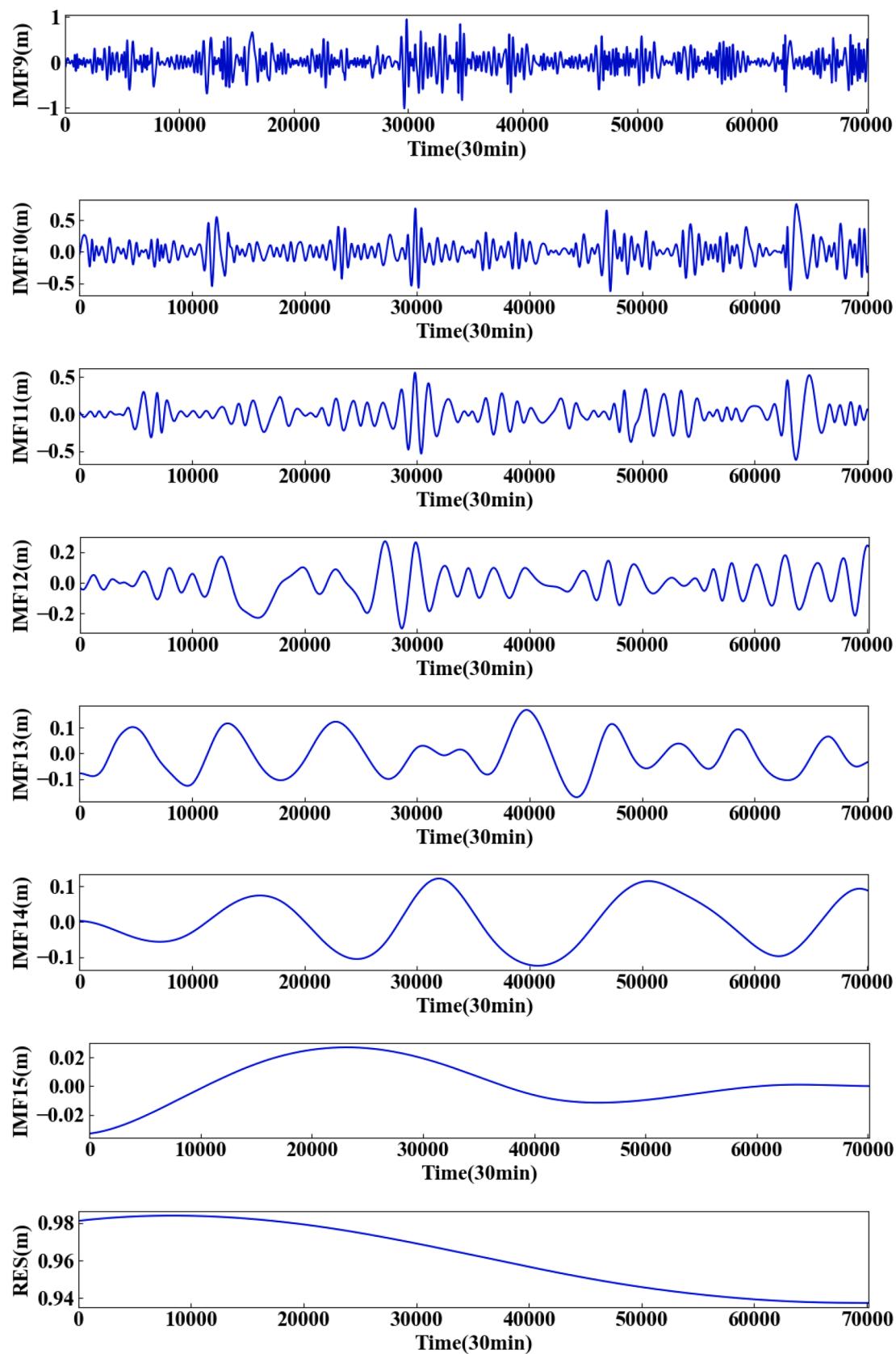


Fig. 6. (continued).

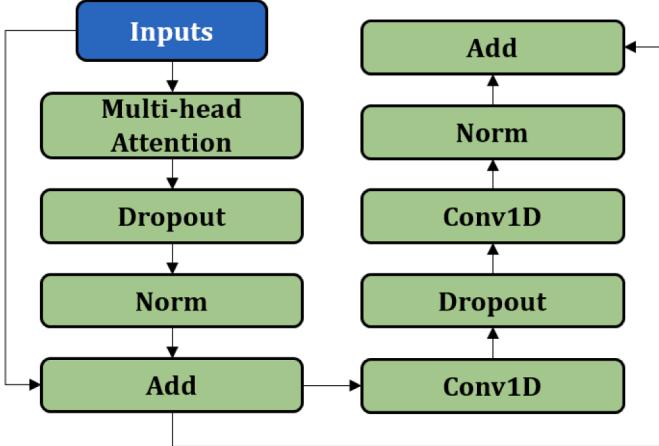


Fig. 7. Structure of a single TF encoder used in current study.

**Table 2**  
Key hyper-parameters for SWH classification model.

Number of heads	Head size	Number of TF blocks	Filter number of 1D-CNN	Kernel size of 1D-CNN
6	6	2	256	1
Number of MLP units	Multi-head dropout	MLP dropout	Number of MLP layer	Batch size
128	0.1	0.4	1	32
General activation	Last activation	Initial learning rate	Objective function	
ReLU	softmax	0.001	Sparse categorical cross-entropy	

rectified linear unit (ReLU) for a fast training to prevent gradient vanishing during backpropagation. The 2-element softmax is used to end up the output layer with a probabilistic categorical result. Based on such an output, taking a single instance as an example, the loss function is defined as follows:

$$CCE = - \sum_{c=0}^{C-1} t_c \log(p_c) \quad (2)$$

where CCE is the abbreviation for categorical cross entropy;  $C (= 2)$  is the number of classification types;  $t_c$  and  $p_c$  are the observation (integer) and probabilistic prediction under the classification type  $c$ , respectively.

The analytical study is performed on Python 3.8 platform based on Keras packages [25]. The two cloud servers for all computation are both equipped with GPU RTX 3070 with 8 G video RAM.

## 4. Results and discussion

The numerical results and multi-aspect discussions are provided in this section through parallel comparisons. The competitors include the EMD and non-EMD models with the mainstream algorithms of TCN, 1D-CNN and LSTM.

### 4.1. Performance criteria

Besides the final average CCE (Formula (2)), the sparse categorical accuracy (CA) is used to evaluate the performance of the predictions from the proposed models. The expression can be drawn as Formula (3):

$$CA = \frac{M_{test}}{N_{test}} \quad (3)$$

where  $M_{test}$  represents the number of predicted labels (binary) that match the true labels in test set; and  $N_{test}$  is the total size of the test set.

### 4.2. Accuracy comparison

The accuracy comparison is discussed in this part. The three parallel models are systematically tuned with key settings as follows:

1. TCN: use two TCN layers with 64 and 32 filters; Dilations equals to [1, 4, 15, 32].
2. 1D-CNN: use three 1D convolutional layers all with filters of 64 and kernel size of 3, and use same padding technique around edges.
3. LSTM: Use two LSTM layers with 32 units; the first LSTM layer returns the complete sequence while the second does not.

#### 4.2.1. Effects of lead time

In this subsection, the accuracy comparisons are conducted under different lead times varying from 6 h to 48 h. To control the variables, a moderate window size of 30 h is used for each experiment and the fixed sampling scheme is adopted to ensure the training and testing datasets unchanged.

The variation for 1-day ahead is shown in Fig. 8 and Table 3, where the accuracy of all EMD model gradually decrease with increasing lead time. The curves of TF and LSTM based model are comparable and nearly overlapped. For short lead time of 6 h, their accuracy on test set is impressively high, reaching 0.991 and 0.988, respectively. They still maintain a high performance around 0.900 when the lead time rises to complete one-day, which shows strong ability in learning far-end related features. However, the TCN and 1D-CNN models present a considerably drop as the lead time enlarges. Both have the initial accuracy over 0.950, but the precision ends up with values around 0.800 at 24 h in advance.

Similarly, Fig. 9 and Table 4 display the plots of variations till two days ahead, showing a significant decrease with increasing inclination. The overall difference between EMD-TF and EMD-LSTM is still micro, and the ending accuracy on 48 h is around 0.720. As before, the TCN and 1D-CNN perform worse, but the absolute difference of TCN accuracy between LSTM and TF becomes narrow while the precision of 1D-CNN is continuously exacerbated by the enlargement of lead time.

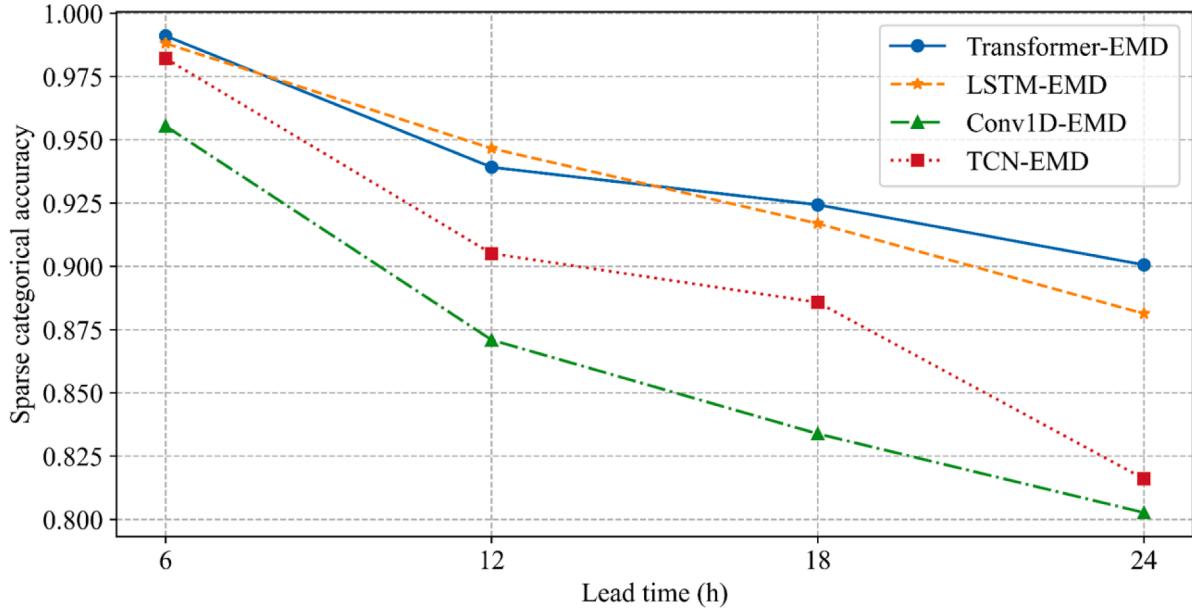
The above analysis indicates that the TF based model can achieve an equivalence even with a subtle higher level of accuracy than LSTM for the task throughout 2-day lead time. The performance is highly promising for the first 24 h but was continuously fading afterwards. At the same time, the exactness of the proposed model can also distinctly surpass the 1D-CNN and TCN based models.

#### 4.2.2. Effects of EMD features

In this subsection, the influences of EMD features are discussed. Here, the models with minimum (6 h) and maximum (48 h) lead time are selected for investigation to exaggerate their difference.

As shown in Fig. 10 and Table 5, for 6 h ahead cases, the EMD features play a pivotal role in decreasing the model loss. The cross-entropies of all models remarkably dive from the platforms with the similar level around 0.210. Among them, even the comparatively worst model 1D-CNN has approximately 50% improvement, the EMD impacts on TF model is the most impressive, reaching nearly 90% of cutdown from the original one.

On the other side, As shown in Fig. 11 and Table 6, for 48 h ahead cases, the influence of EMD features is inconspicuous, and the differences between EMD and non-EMD models are within 25%. It should be noticed that the loss of non-EMD TF is even not the lowest among all models. However, when EMD features are included, an obvious improvement for TF is obtained compared with the counterparts. Such a phenomenon may attribute to that the multi-head attention layers can better extract effective information from IMF series for future evaluation. Therefore, the EMD features are more appropriate workmates with TF than with others.



**Fig. 8.** Model accuracy changes w.r.t. different lead time from 6 to 24 h.

**Table 3**  
Accuracy variations of EMD parallel models for lead time within 1day.

	6h	12h	18h	24h
EMD-TF	0.991	0.939	0.924	0.900
EMD-Conv1D	0.955	0.870	0.833	0.802
EMD-TCN	0.982	0.905	0.885	0.816
EMD-LSTM	0.988	0.946	0.916	0.881

#### 4.2.3. Effects of window size

Besides, the window size, such as multi-step length in some publications, may be influential too. Therefore, a systematical analysis of the loss under each lead time with various window sizes is conducted in this subsection.

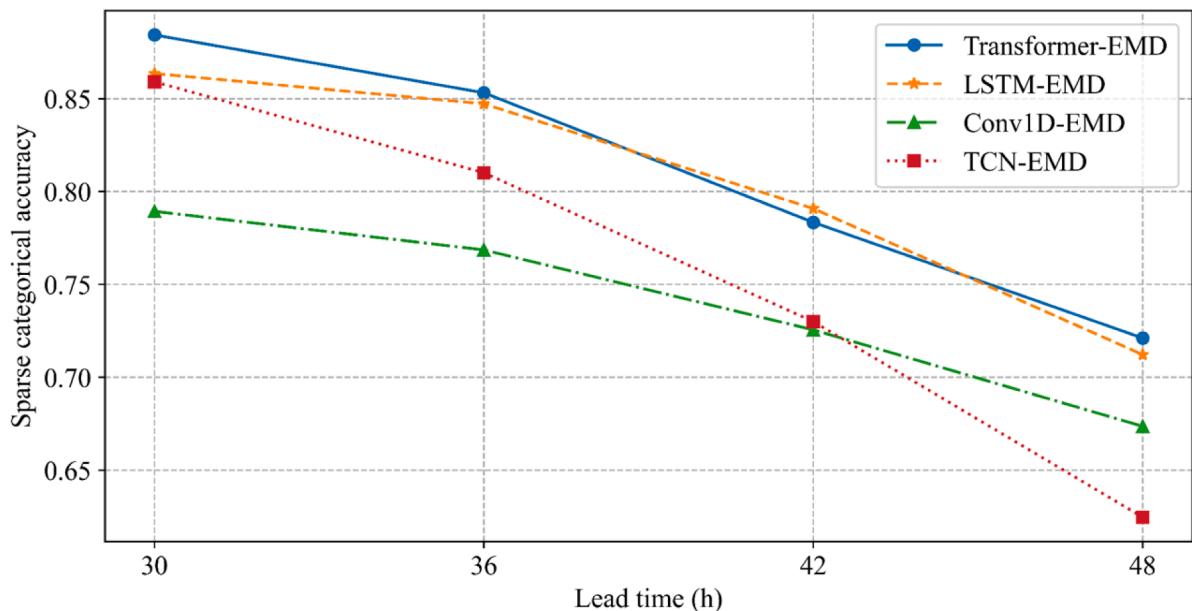
As shown in Fig. 12, to examine the ability of generalization, various sampling datasets were used under fixed lead time and fixed window

size (solid lines), while the dots represent their mean loss. Then, the dashed lines indicate the tendency of the changes in cross-entropy for test sets as the window size is increasing.

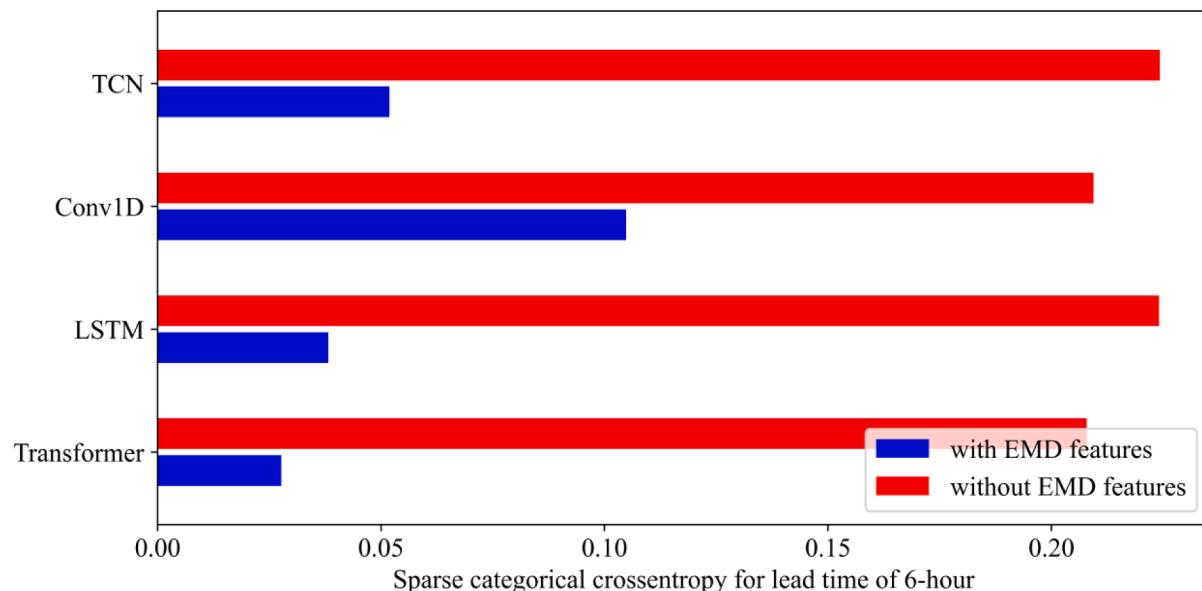
As it can be found from Fig. 12, when lead time is increased from 6 h to 24 h, the variation amplitude of the loss among different window sizes becomes larger. Also, the ranges of variation in different samplings are also increased as the lead time becomes longer. At the same time, the

**Table 4**  
Accuracy variations of EMD parallel models for lead time from 1 to 2days.

	30h	36h	42h	48h
EMD-TF	0.884	0.853	0.783	0.721
EMD-Conv1D	0.789	0.768	0.725	0.673
EMD-TCN	0.85905	0.810089	0.72997	0.624
EMD-LSTM	0.863	0.847	0.790	0.712



**Fig. 9.** Model accuracy changes w.r.t. different lead time from 30 to 48 h.



**Fig. 10.** Loss chart comparisons for different models with or without EMD features under a lead time of 6 h.

**Table 5**

Loss value comparisons for different models with or without EMD features under a lead time of 6 h.

	With EMD features	Without EMD features	Percentage
TF based	0.0277	0.207	13.38%
Conv1D based	0.1048	0.209	50.14%
TCN based	0.0519	0.224	23.17%
LSTM based	0.0382	0.224	17.05%

best window size is gradually right shifted: 12 h for lead times of 6 h; 18 h for lead time of 12 h; 24 h for lead time of both 18 and 24 h. This may be partly because that larger lead time requires more historical information to build more long-term connections. However, for short lead time cases, the far-step information may be redundant and thus aggravated the training burden and reduced the accuracy.

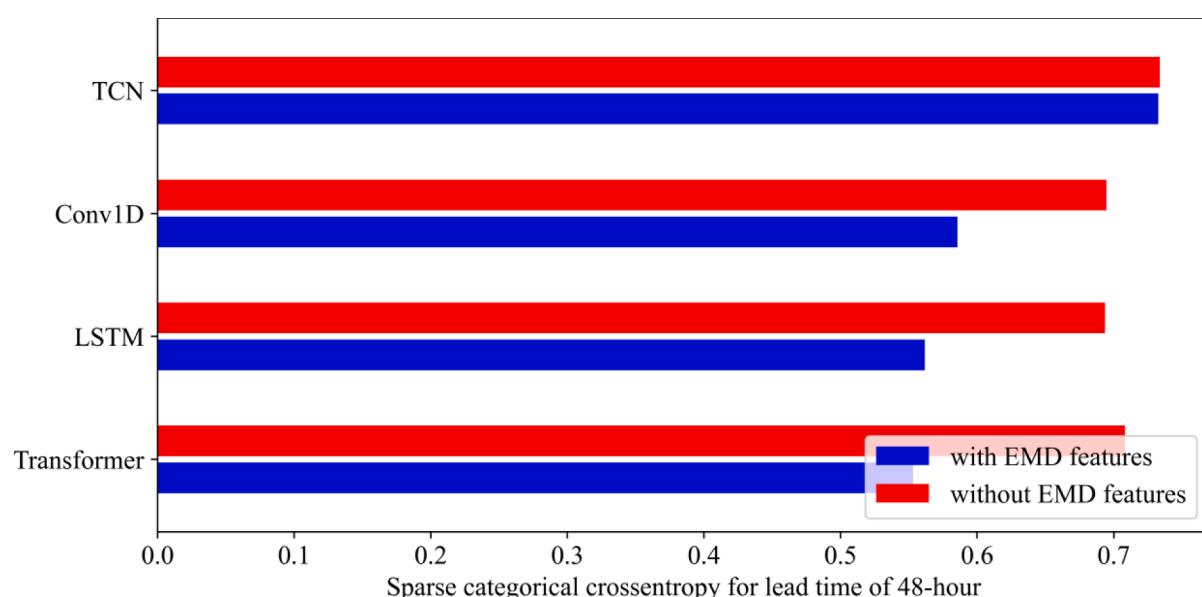
However, the patterns are becoming different for further prolonging lead time, as shown in Fig. 13. For lead time of 30 h, the rules are

inherited, and a larger amplitude of fluctuation can be observed among different window sizes. In addition, the best window size keeps on moving to 30 h as the yellow dots indicate. However, such a pattern stops when lead time equals to 36 h. From 36 h to 48 h, the three curves are generally monotonously climbing as the window size increases, and the best window size locates at the minimum step length of 6 h. A probable explanation for this phenomenon could be that most IMF features focus on the local extreme features. When lead time becomes

**Table 6**

Loss value comparisons for different models with or without EMD features under a lead time of 48 h.

	With EMD features	Without EMD features	Loss percentage
TF based	0.553	0.708	78.11%
Conv1D based	0.585	0.694	84.3%
TCN based	0.732	0.733	99.8%
LSTM based	0.561	0.693	80.95%



**Fig. 11.** Loss chart comparisons for different models with or without EMD features under a lead time of 48 h.

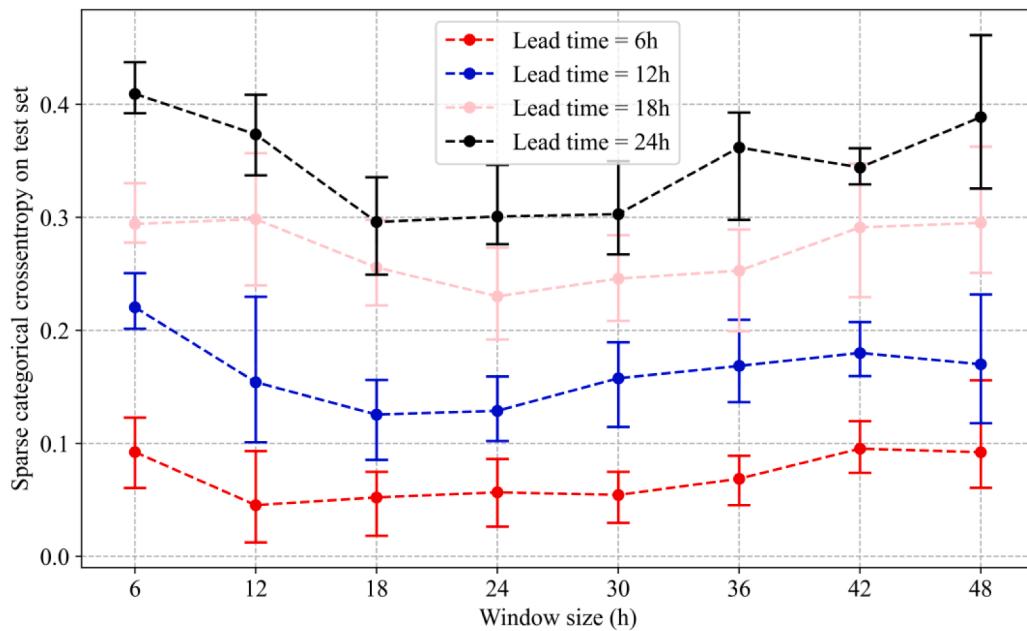


Fig. 12. Test loss variation with different window sizes under the lead time changing from 6 h to 24 h.

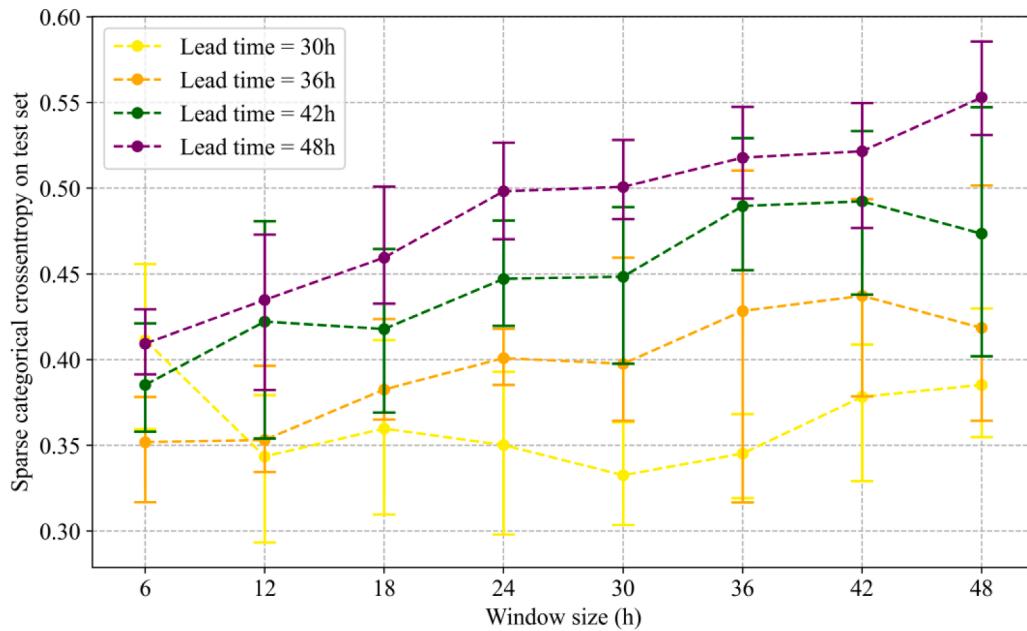


Fig. 13. Test loss variation with different window sizes under the lead time changing from 30 h to 48 h.

sufficiently long, the rules reflecting from those features, such as the patterns of local noise development, are meaningless. Therefore, as the window size becomes longer, the noneffective input can only make the model hard to train and more likely to be overfitting.

#### 4.3. Efficiency comparison

Apart from the accuracy, the efficiency is also significant in evaluating the model, especially for real dynamic applications. In this work, the representative 30(lead time)/30(window size) cases are used to evaluate the time cost of different EMD models. Due to the different sampling datasets, the averaged total computational time and the number of epochs to obtain the best model through early-stop technique are shown in Table 7.

Table 7

Averaged total time cost and epoch number comparison between different EMD models.

	Averaged time cost	Averaged epoch	Time cost ratio
EMD-TF	76.221 s	15	1.000
EMD-TCN	84.820 s	7	1.113
EMD-Conv1D	35.668 s	5	0.468
EMD-LSTM	905.426 s	24	11.879

As it can be seen from Table 7, the total time cost of TF based model ranks second out of four models, which cost about double time than the first place EMD-Conv1D. The drawbacks EMD-LSTM are apparent. On the one hand, though all four models have acceptable number of

convergence steps, the EMD-LSTM still needs comparatively more epochs. On the other hand, noticeably, the total time cost of EMD-LSTM is considerably larger than any other models and over tenfold of EMD-TF for its essence of serial structure. Therefore, considering the accuracy discussed in Section 4.2, the EMD-TF appeals to be a promising and high-efficient model for the current task.

In order to apply the current model in real-world applications, it is necessary to establish both an offline and an online system that work collaboratively. The online system will obtain historical measurements from sensors, use the model in-hand to make classifications, and continuously transmit new data to the offline system. Meanwhile, the offline system will periodically update the model based on pre-defined error thresholds and send the updated model back to the online system for deployment. By setting up this coordinated system, the model can be optimized and refined over time, providing increasingly accurate and reliable predictions for a variety of SWH classification tasks.

## 5. Conclusion

In this study, a new framework, empirical mode decomposition with Transformer encoder (EMD-TF), is proposed to improve wave energy management by classifying the level of significant wave height in advance. The high-level and low-level thresholds of the significant wave height (SWH) are set to be 2.0 m and 1.5 m in this study. Through numerical studies, the model presents the following impressive advantages:

- 1 The current model can provide up to 9.80% and 9.97% higher accuracy compared with temporal convolutional network and one-dimensional convolutional neural network based models within lead time of 24 h and 48 h. The accuracy of long short-term memory network combine with empirical mode decomposition (EMD-LSTM) is comparable with a subtle less than the current model.
- 2 For a short lead time of 6 h, the features from empirical mode decomposition can largely reduce the testing loss for all parallel models but become the most effective on Transformer models, thus cutting down nearly 90% of the categorical cross entropy.
- 3 For a long lead time up to 48 h, the influence of features from empirical mode decomposition is largely weaken but is still more obvious on Transformer.
- 4 For current model, distinct patterns are shown regarding the effects of window size for different ranges of lead times. The best window size for lead time less than 30 h displaying a right-shift tendency while the minimal window size is the best for lead time over 30 h.
- 5 Efficiency tests indicate that the time cost of EMD-LSTM is over tenfold than that of the EMD-TF model, suggesting that the current framework could be a more high-efficient alternative in SWH classification tasks.

It is important to acknowledge that the current data source is restricted to the North Atlantic Ocean. However, we anticipate that the proposed method has the potential to be adapted to various datasets from different regions worldwide. This can be achieved by modifying the high-level and low-level SWH thresholds to match the local environmental conditions. While the patterns of model performance with lead time may vary when applied to different regions, we believe that the fundamental tendency and effectiveness of the model will remain consistent.

In the future, further research is needed to fully understand frequency or phase [26] characteristics and their neural network supports, such as weight distribution, in order to form more robust explanations or theories.

## 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.

## Data availability

Data will be made available on request.

## Acknowledgements

The financial support from the National Natural Science Foundation of China (No. 61973208) is gratefully acknowledged. Also, the paper is sponsored by “Shuguang Program” (18SG36) that is supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission.

## References

- [1] Caloiero T, Aristodemo F, Ferraro DA. Annual and seasonal trend detection of significant wave height, energy period and wave power in the Mediterranean Sea. *Ocean Eng* 2022;243:110322.
- [2] Gomez-Orellana AM, Guijo-Rubio D, Gutiérrez PA, Hervás-Martínez C. Simultaneous short-term significant wave height and energy flux prediction using zonal multi-task evolutionary artificial neural networks. *Renew Energy* 2022;184: 975–89.
- [3] Abdulla FAR, Ningsih NS, Al-Khan TM. Significant wave height forecasting using long short-term memory neural network in Indonesian waters. *J Ocean Eng Mar Energy* 2022;8(2):183–92.
- [4] Littardi A, Hildeman A, Nicolaou MA. Deep learning on the sphere for multi-model ensembling of significant wave height. In: Proceedings of the ICASSP 2022–2022 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2022. p. 3828–32.
- [5] Lou R, Lv Z, Guizani M. Wave height prediction suitable for maritime transportation based on Green Ocean of things. *IEEE Trans Artif Intell* 2022.
- [6] Yang S, Deng Z, Li X, Zheng C, Xi L, Zhuang J, Zhang Z. A novel hybrid model based on STL decomposition and one-dimensional convolutional neural networks with positional encoding for significant wave height forecast. *Renew Energy* 2021;173: 531–43.
- [7] Feng Z, Hu P, Li S, Mo D. Prediction of significant wave height in offshore china based on the machine learning method. *J Mar Sci Eng* 2022;10(6):836.
- [8] Hao W, Sun X, Wang C, Chen H, Huang L. A hybrid EMD-LSTM model for non-stationary wave prediction in offshore China. *Ocean Eng* 2022;246:110566.
- [9] Chen Y, Wang Y, Dong Z, Su J, Han Z, Zhou D, Bao Y. 2-D regional short-term wind speed forecast based on CNN-LSTM deep learning model. *Energy Convers Manag* 2021;244:114451.
- [10] Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. *Energy* 2018;148:461–8.
- [11] Li T, Hua M, Wu XU. A hybrid CNN-LSTM model for forecasting particulate matter (PM<sub>2.5</sub>). *IEEE Access* 2020;8:26933–40.
- [12] Shahid F, Zameer A, Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos Solit Fract* 2020;140:110212.
- [13] Yu X, Liu Y, Sun Z, Qin P. Wavelet-based Resnet: a deep-learning model for prediction of significant wave height. *IEEE Access*; 2022.
- [14] Zhou S, Bethel BJ, Sun W, Zhao Y, Xie W, Dong C. Improving significant wave height forecasts using a joint empirical mode decomposition-long short-term memory network. *J Mar Sci Eng* 2021;9(7):744.
- [15] Huang W, Dong S. Improved short-term prediction of significant wave height by decomposing deterministic and stochastic components. *Renew Energy* 2021;177: 743–58.
- [16] Chen Y, Dong Z, Wang Y, Su J, Han Z, Zhou D, Bao Y. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. *Energy Convers Manag* 2021;227:113559.
- [17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I. Attention is all you need. *Adv Neural Inf Process Syst* 2017:30.
- [18] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Liu HH. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In: Proceedings of the royal society of London. Series A: mathematical, physical and engineering sciences. 454; 1998. p. 903–95.
- [19] Pokhrel P, Ioup E, Simeonov J, Hoque MT, Abdelguerfi M. A transformer-based regression scheme for forecasting significant wave heights in oceans. *IEEE J Ocean Eng* 2022;47(4):1010–23.
- [20] Putri DA, Adytia D. Time series wave forecasting with transformer model, case study in Pelabuhan Ratu, Indonesia. In: Proceedings of the 10th international conference on information and communication technology (ICoICT). IEEE; 2022. p. 430–4.
- [21] Behrens J, Timmerman R, Terrill E, Merrifield S, Jensen RE. CDIP observations of recent extreme wave conditions on US coasts. *Shore Beach* 2021;89(2):41.
- [22] Leatherman SP, Douglas BC, Kearney. In: *Sea level rise: history and consequences*. Academic; 2000.

- [23] Liu X, Huang W, Gill EW. Estimation of significant wave height from X-band marine radar images based on ensemble empirical mode decomposition. *IEEE Geosci Remote Sens Lett* 2017;14(10):1740–4.
- [24] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [25] Gullu A, Pal S. Deep learning with keras. Packt Publishing Ltd; 2017.
- [26] Xu X, Hu S, Shi P, Shao H, Li R, Li Z. Natural phase space reconstruction-based broad learning system for short-term wind speed prediction: Case studies of an offshore wind farm. *Energy* 2023;262:125342.