ORIGINAL RESEARCH **Open access**

Transformer-Based Significant Wave Height Classification Using Empirical Mode Decomposition for Wave Energy Management

David Thompson[1*], Sarah Mitchell[1], Rachel Adams[2]

Abstract

Forecasting significant wave height (SWH) plays a critical role in wave energy management. Although numerous machine learning investigations have concentrated on precisely forecasting specific SWH values a few hours ahead, practical real-world applications fundamentally prioritize determining the qualitative wave height level. This manuscript introduces a time-series classification framework for SWH, constructed upon a Transformer encoder (TF) and empirical mode decomposition (EMD). This system offers prediction lead times ranging from 6 to 48 h, utilizing predefined fixed thresholds of 2 m for elevated wave conditions and 1.5 m for lower wave conditions. The efficacy of the proposed method is evaluated against three standard algorithms, both with and without EMD features. Analysis of datasets derived from Atlantic Ocean buoy observations reveals an optimal mean accuracy of 99.1% at a 6 h lead time, alongside an average training duration of 75 s. These findings highlight both the precision and computational efficiency of the introduced model. Consequently, this research delivers practical methodologies and foundational references for operational SWH forecasting systems.

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

*Correspondence:
David Thompson
david.thompson@utoronto.ca
[1] Department of AI-Based Load Forecasting and Demand Response, Faculty of Engineering, University of Toronto, Toronto, Canada
[2] Department of Intelligent Electrical Energy Systems, Faculty of Applied Science, University of British Columbia, Vancouver, Canada

Introduction

Marine waves represent a prominent category of renewable energy resources, capable of being transformed into electrical power via attenuators alongside other specialized mechanisms. The subsurface power flow intensities of these waves range between 0.1 and 0.3 kW/m² and 2–3 kW/m², which is approximately an order of magnitude greater than surface wind energy. Furthermore, oceanic wave energy exhibits superior reliability compared to wind power [1]. Despite wave energy technology remaining in a developmental phase, it holds substantial promise for prospective energy utilization.

In the context of ocean waves, significant wave heights (SWHs) serve as fundamentally critical parameters of the marine environment [2]. Consequently, comprehending and accurately forecasting future SWH conditions is imperative for the effective administration of wave energy operations. Specifically, during mechanized energy conversion, it is necessary to possess the operational capacity to suspend equipment functionality when wave energy resources are inadequate or when conditions exceed the hardware's operational limits [1]. The efficient extraction and application of ocean energy

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

inherently rely on theoretical modeling and simulation. Conventional numerical simulation techniques, including third-generation wave models (e.g., SWAN), demand substantial computational resources to process extensive oceanographic datasets and resolve Navier-Stokes equations. Consequently, these traditional models are inadequate for short-term forecasting during wave evolution, particularly for predictive horizons of only a few hours [3].

In recent years, methodologies rooted in statistics and machine learning have emerged as viable solutions to this forecasting challenge [4]. The fundamental premise of these approaches is to capture nonlinear relationships across sequential time steps using historical time-series data. To date, a diverse array of time-series architectures has been proposed, encompassing temporal convolutional networks (TCN) [5], 1D convolutional neural networks (CNN) [6], and, predominantly, configurations based on recurrent neural networks (RNN) [7]. Typically, convolutional network-based architectures exhibit distinct advantages regarding computational time budgets. In contrast, RNN-based frameworks—such as long short-term memory (LSTM) networks or gated recurrent units (GRU)—offer enhanced versatility and demonstrate superior predictive accuracy across diverse datasets [8]. Analogous methodologies have also been successfully implemented in adjacent domains, such as wind velocity forecasting [9], solar power estimation [10], atmospheric pollution tracking [11], and predictive modeling of the COVID-19 pandemic [12].

Broadly speaking, determining whether the exact SWH is 2.2 m or 2.4 m is often inconsequential for operational decision-making within wave energy management. The pivotal consideration remains identifying the opportunity for significant energy extraction during periods of a “high wave level.” Conversely, predicting SWH values of 0.7 m versus 0.9 m elicits identical operational responses, as both signify a “low” sea state condition. Therefore, through the application of machine learning techniques, this predictive challenge can be re-framed from a regression task into a classification endeavor, continuously employing historical time-series data as the primary input. This paradigm shift facilitates the generation of more precise and extended-range forecasts. Regression models often prove less effective in these specific dimensions, primarily because the computational exertion required to pinpoint exact numerical values may be redundant. Nevertheless, existing academic literature exploring this classification-based approach remains notably sparse.

When executing time-series classification (TSC) for SWH, predictive accuracy is accompanied by at least two other critical considerations. The primary concern is the maximum prediction horizon over which the model maintains adequate precision. These two objectives often conflict; specifically, the model’s ability to preserve accuracy typically degrades as the lead time extends. The secondary concern involves the computational expense of training, which dictates the feasibility of implementing dynamic, real-world model updates. Furthermore, in SWH-related operations where transmitting observational records is challenging—particularly during turbulent sea states—minimal training overhead is highly advantageous for enabling edge computing via Internet of Things (IoT) hardware or localized edge servers.

Across widely used architectures, including diverse CNN- and RNN-based frameworks [5], it is generally infeasible to achieve an optimal balance between accuracy and computational time cost by relying solely on a single paradigm. For example, CNN architectures inherently struggle to establish long-range temporal dependencies, whereas RNN architectures necessitate exorbitant training budgets. Consequently, the Transformer (TF) architecture emerges as a highly promising alternative, distinguished by its self-attention mechanisms and robust parallel processing capabilities.

Additionally, preceding research indicates that time-series decomposition techniques may address the aforementioned limitations. In the investigation conducted by Yang *et al*. [6], seasonal and remainder components were extracted utilizing the Seasonal and Trend decomposition using Loess (STL) technique. This method demonstrated superior performance in shallow-water environments, particularly for lead times exceeding 24 h. Relatedly, independent research has shown that STL decomposition requires negligible processing time, taking less than 0.08 s. In a separate study [13], wavelet transformation was integrated to enhance a ResNet model, simultaneously decreasing the mean squared error by up to 18% and curtailing training expenditures. The most widely adopted decomposition strategy is a strictly a posteriori approach: empirical mode decomposition (EMD). Multiple investigations have coupled EMD with various machine learning algorithms to substantially improve predictive precision [8]. While prior studies have highlighted its advantageous properties (including its a posteriori nature [14] and self-adaptive capabilities [15]), the conventional procedure of initially

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

decomposing modes, individually forecasting each component, and subsequently aggregating the results (abbreviated herein as the DPS framework) proves overly intricate for handling massive datasets and facilitating real-time updates [16]. Crucially, this traditional DPS approach lacks direct applicability to the current SWH classification objective, necessitating the adoption of a more appropriate operational framework.

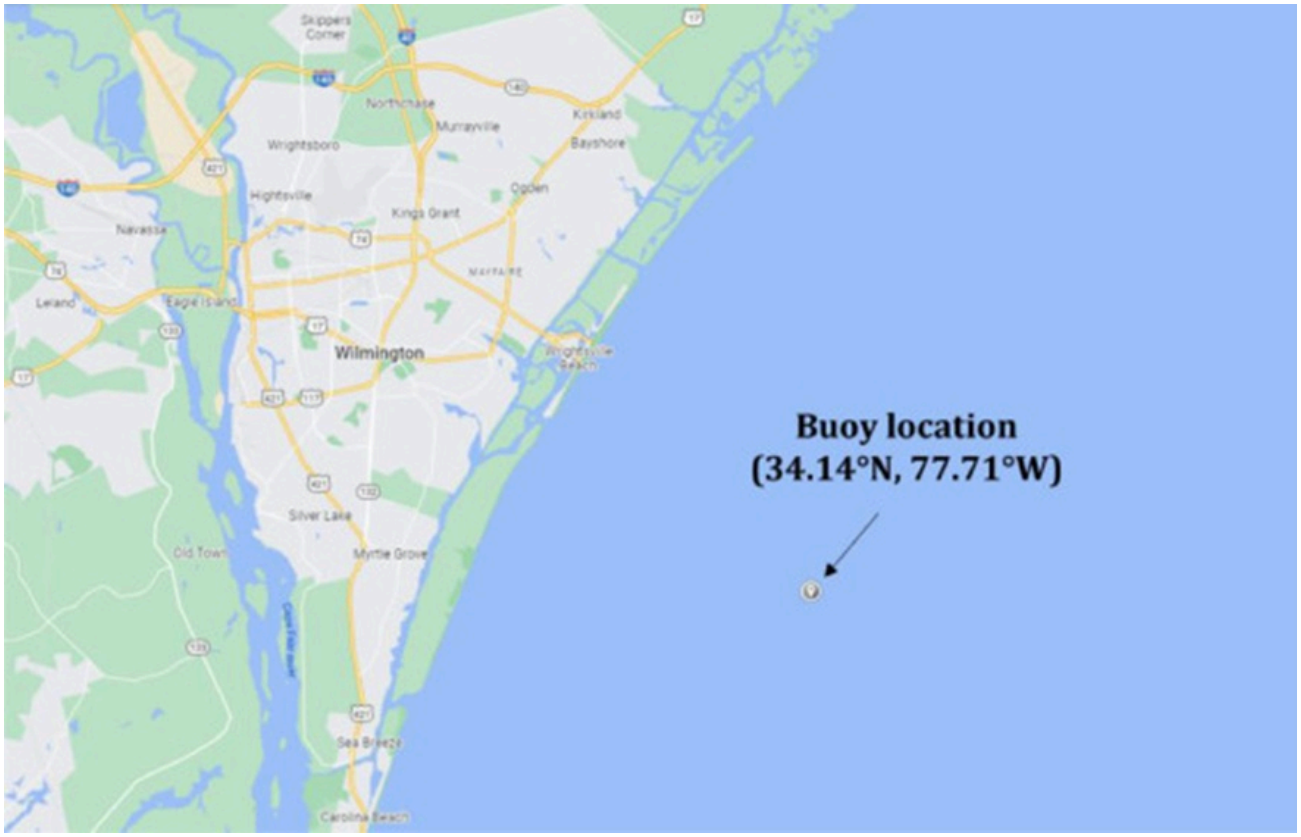
Accordingly, this research introduces a novel TSC architecture for SWH forecasting, integrating the Transformer encoder (TF) [17] with a robust signal decomposition technique, specifically empirical mode decomposition (EMD) [18]. A select number of recent investigations by Pokhrel *et al*. [19] and Putri and Adytia [20] have tentatively demonstrated the proficiency of standalone TF models in SWH prediction contexts, applied either to resolving differential equations [19] or managing extended lead times in contrast to short-term scenarios [20]. Building upon methodologies established in our preceding publication [16], the present study directly employs EMD-derived sub-series as input features to augment the TF model's performance for forecasting horizons between 6 and 48 h in advance. This architectural design offers a more streamlined and highly effective paradigm for constructing models explicitly tailored to time-series classification objectives. An in-depth evaluation of the model's performance is conducted via parallel comparisons focusing on predictive accuracy, robustness, and computational time efficiency.

Dataset

This section outlines the origins of the SWH measurements alongside the applied pre-processing methodologies.

Data source

The foundational historical SWH data were retrieved from the prominent Coastal Data Information Program (https://cdip.ucsd.edu/) [21], an initiative established in 1975 with financial support from the US Army Corps of Engineers. As depicted in **Figure 1**, the offshore buoy utilized in this study is situated in the Masonboro Inlet within the North Atlantic Ocean (34.14° N, 77.71° W), positioned at a water depth of 15.700 m.



© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Figure 1. Buoy location for SWH measurements in Masonboro Inlet near Wilmington [21].

As presented in **Table 1**, the extracted dataset spans four years from 2017 to 2020, featuring a temporal resolution of 30 min. The aggregation yields more than 70,000 discrete timestamps, with a global maximum SWH of 5.100 m (level-6 sea state) and an absolute minimum of 0.260 m. Furthermore, both the median and mean values approximate 0.900 m, accompanied by a moderate standard deviation of 0.414 m.

Table 1. The statistical information of raw SWH buoy data from CDIP [21].

Metric	Value
Total data points	70,124
Sampling interval (min)	30
Average value (m)	0.945
Middle value (m/s)	0.850
Highest recorded value (m)	5.100
Lowest recorded value (m)	0.260
Standard deviation (m)	0.414

The unprocessed observations are remarkably comprehensive, containing a mere 4 missing entries. To rectify these omissions and finalize the dataset, linear interpolation was applied; the generated continuous curve is displayed in **Figure 2**.

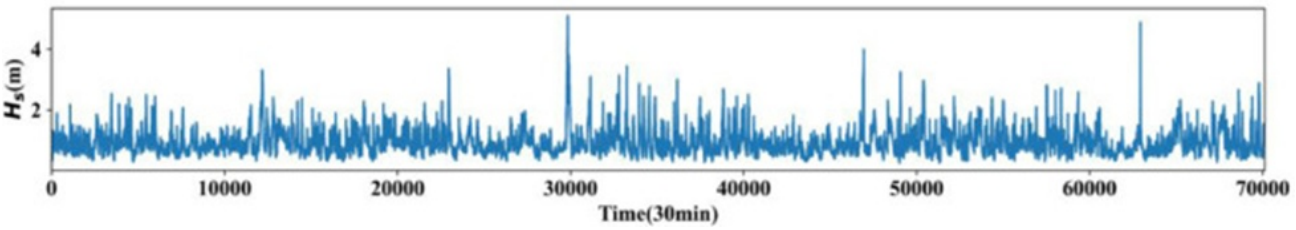


Figure 2. The complete SWH dataset from CDIP (2017–2020) [21].

Data pre-processing

Time-series classification (TSC) architectures demand rigorous data conditioning. Consequently, this subsection outlines the critical operational definitions for the targeted problem alongside the methodology for dataset construction.

Key parameters

Initially, this segment establishes four fundamental parameters relevant to the analytical task: window size, lead time, the high-level SWH threshold, and the low-level SWH threshold. A visual representation of these variables is provided in **Figure 3.**

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

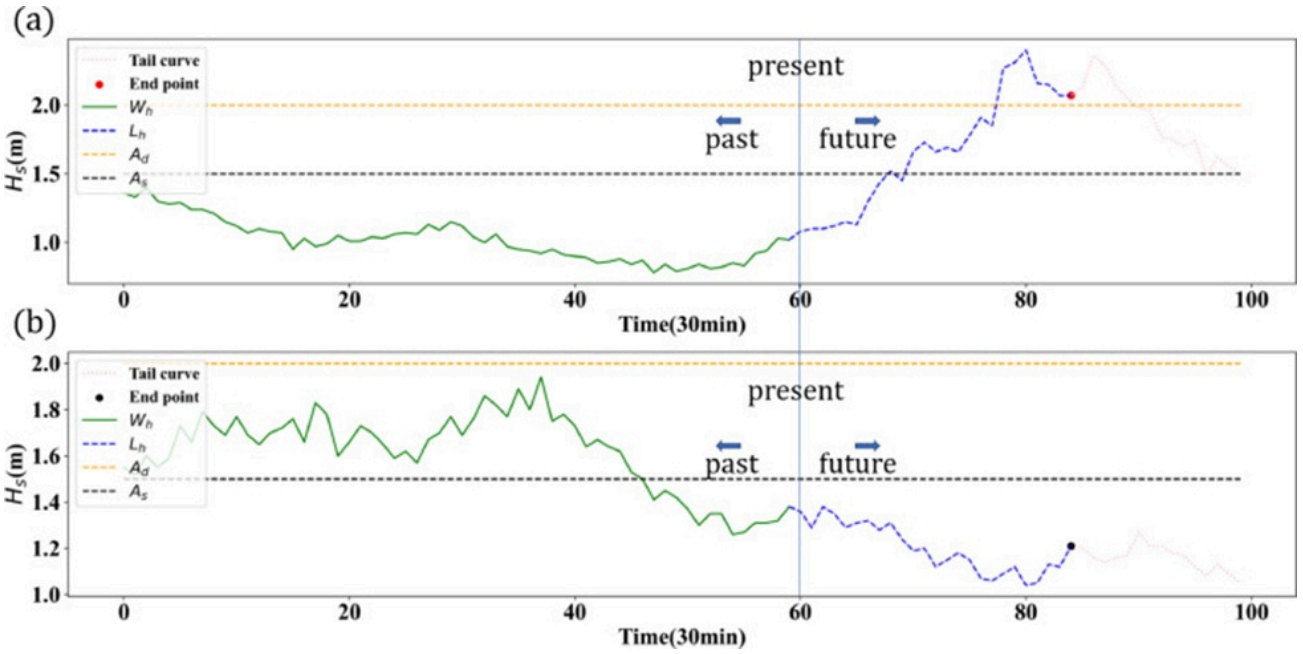


Figure 3. The illustration of 4 key parameters for SWH classification: the green solid lines are known information, and the red/black color of the endpoint represents the type of the wave level: (a) High-level ending, and (b) low-level ending.

As illustrated in **Figure 3**, the solid green line spans a duration of Wh, representing the sequence where all observational data points are confirmed. The dashed blue segment, extending over Lh, denotes the future horizon where data points remain unknown and unforecasted. The qualitative categorization (rather than the exact numerical value) of the terminal point is determined as follows: an endpoint projected to meet or exceed Ad (the horizontal dashed orange line) is designated as a “high-level” event and highlighted in red **(Figure 3a).** Conversely, an endpoint forecasted to fall below As (the horizontal dashed black line) is classified as a “low-level” event and highlighted in black **(Figure 3b)**. Within this illustrative example, Wh and Lh are set to 30 h and 12 h, respectively; the specific impacts of these durations will be analyzed in detail. The threshold values Ad = 2 m and As = 1.5 m remain constant throughout this research, generally corresponding to the demarcations for medium/large waves and small waves as delineated in Douglas *et al*. [22].

Dataset generation

Utilizing the fully interpolated measurements in conjunction with the aforementioned parameters, this subsection delineates the methodology for constructing both the training and testing datasets. The complete workflow is visually summarized in **Figure 4.**

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

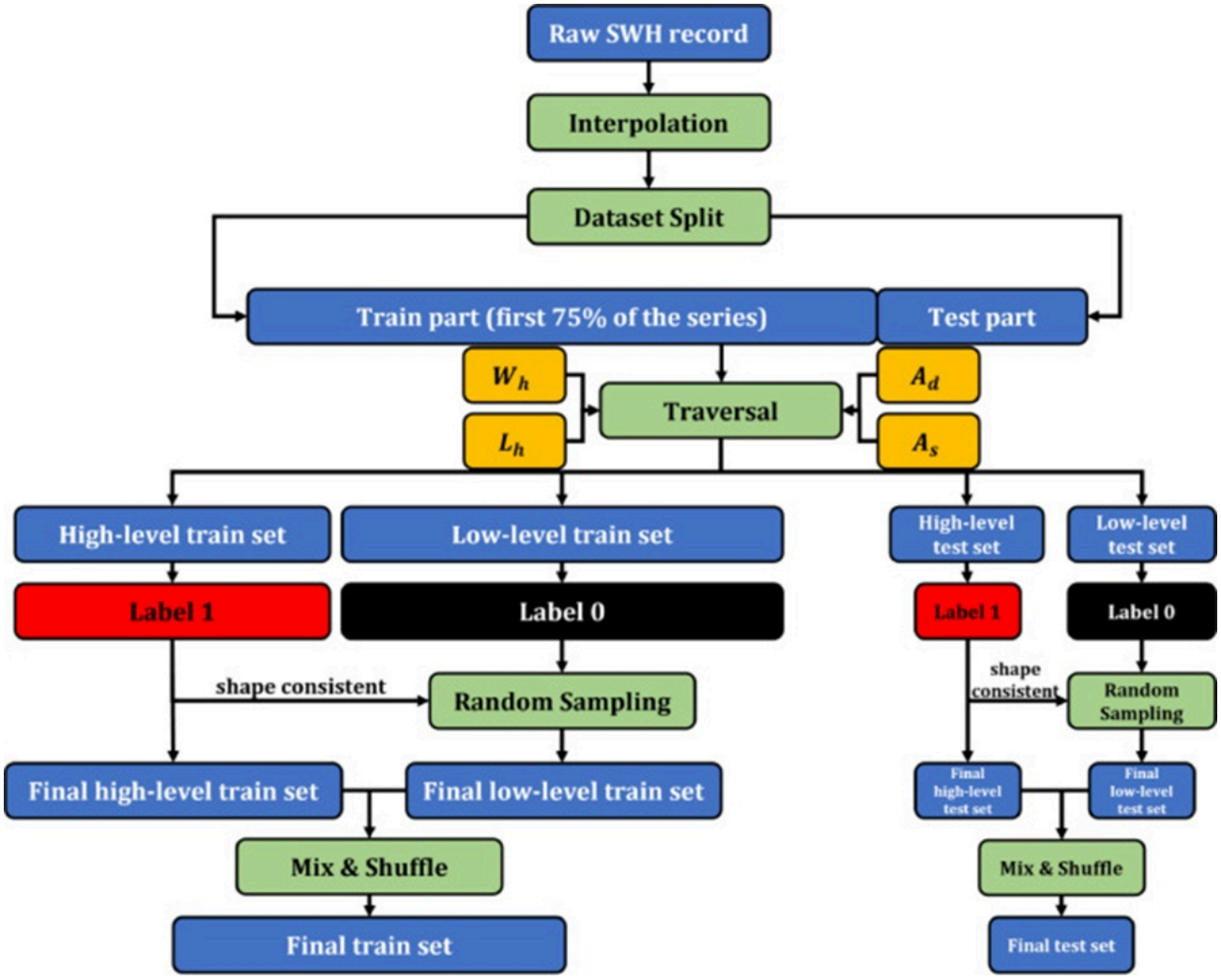


Figure 4. The flowchart of dataset preparation for machine learning models.

The continuous time series is chronologically partitioned, allocating the initial 75% to the training subset and the remaining 25% to the testing subset. Following this, both subsets are independently processed utilizing the four primary parameters, yielding four distinct categorical groups: high-level train, high-level test, low-level train, and low-level test. The high-level instances are assigned a binary label of 1, whereas the low-level instances are assigned a label of 0. Subsequently, to guarantee symmetrical model training, the dataset is balanced by down-sampling the more abundant low-level instances to match the frequency of the high-level instances. Finally, these consolidated high-level and low-level aggregates are merged and randomized before being ingested by the machine learning algorithms.

Materials and Methods

This research employs a deep learning architecture that integrates a feature extraction mechanism powered by EMD alongside TF encoders, culminating in a fully connected dense layer to produce the final classification. The comprehensive architectural framework is depicted in **Figure 5.**

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

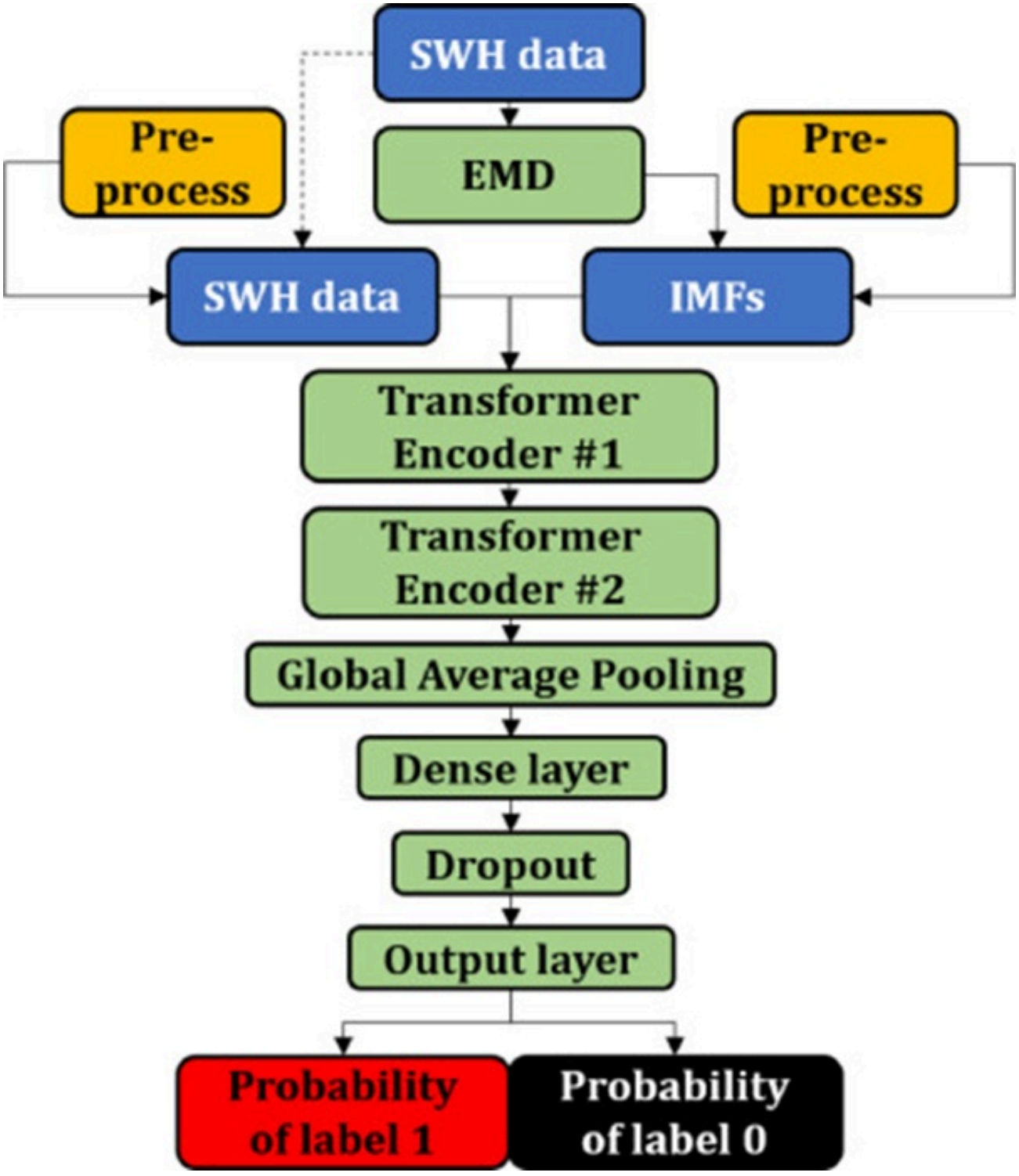


Figure 5. The flowchart of the machine learning model for SWH prediction.

The model's structural design is highly streamlined. Initially, the SWH sequence undergoes signal decomposition via the EMD technique, yielding multiple intrinsic mode functions (IMFs). Following this, the original SWH data and these derived sub-components are concatenated and subjected to the pre-processing protocols. The formalized training arrays are sequentially processed through two distinct TF encoders to capture the complex, nonlinear correlations bridging the extracted features and discrete time steps. A global average pooling operation is subsequently applied to compress the

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

tensor dimensions into a two-dimensional format, preparing it for the dense network. Ultimately, two fully connected dense layers—comprising one hidden layer and one output layer—are appended to generate a probability matrix indicating the predicted wave state. In contrast, a dropout layer is incorporated to enhance the model's generalized robustness.

Empirical mode decomposition

The EMD algorithm, originally introduced by Huang *et al*. [18], has been widely implemented over subsequent decades due to its inherent superiority in adaptively processing nonlinear and non-stationary signals [23].

The sequential sub-components isolated via EMD are termed intrinsic mode functions (IMFs), which adhere to two fundamental conditions: (1) the absolute discrepancy between the total number of local extrema and zero-crossings must not exceed one; and (2) the local mean value, defined by the envelope of local maxima and local minima, must be strictly zero at every individual timestamp. As illustrated in **Figure 6**, the decomposition in this specific study yields a total of 15 IMFs, where the terminal sequence functions as the overall residual term. The linear summation of these individual components perfectly reconstructs the original signal.

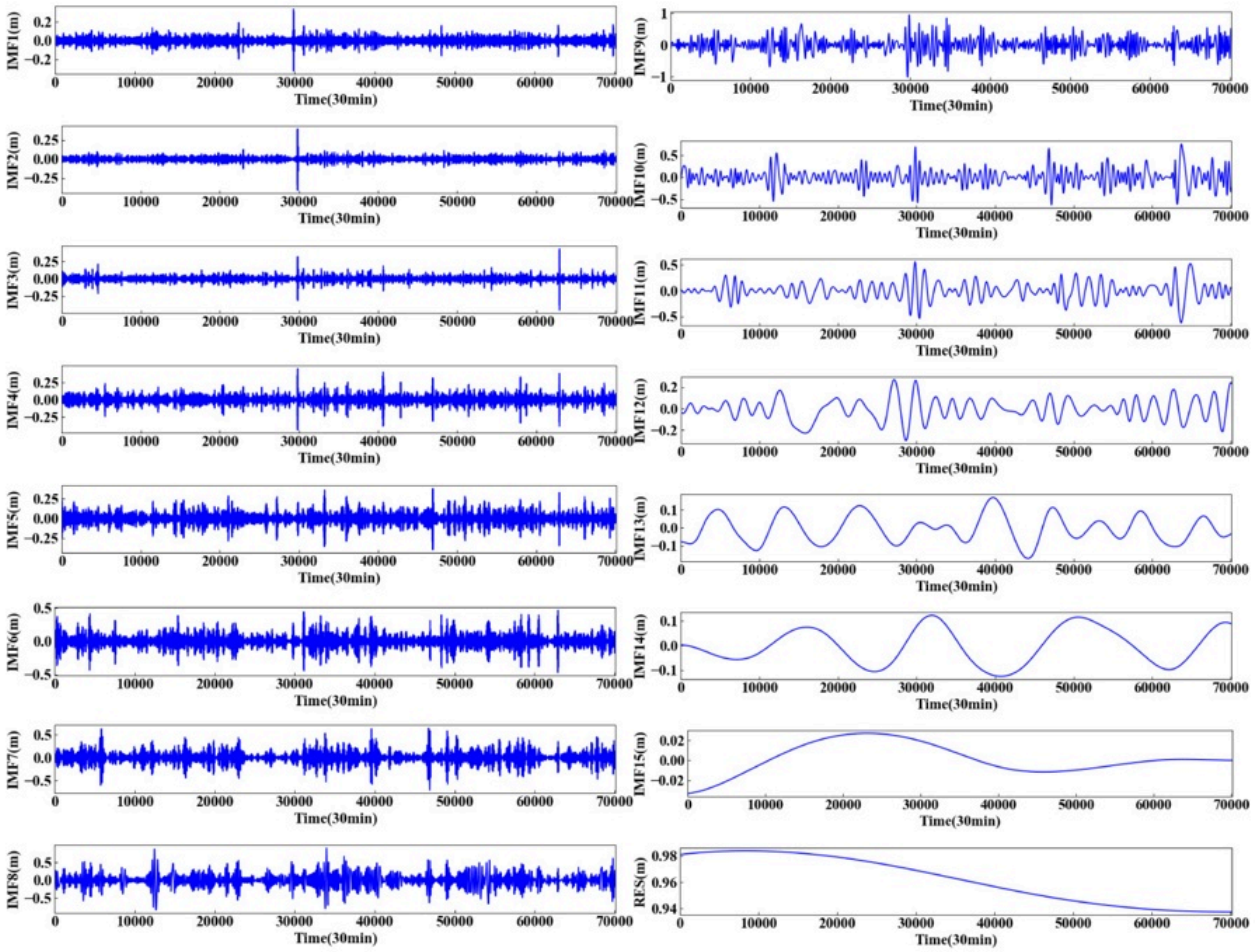


Figure 6. All 15 intrinsic mode functions and the residue of the SWH series after EMD processing.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

When observing **Figure 6** in comparison to the unprocessed data sequence in **Figure 2**, the IMFs exhibit significantly enhanced regularity. This is partially attributable to the aforementioned mathematical constraints, which effectively isolate and remove long-term macro-fluctuations within individual components, thereby stabilizing each IMF. Furthermore, as the decomposition progresses systematically from high-frequency to low-frequency domains, the intricate, localized stochastic variations that drastically dictate prediction precision [16] are effectively decoupled from the bulk data. This strategic extraction fundamentally elevates the subsequent TSC classification metrics.

TF encoder

This subsection succinctly outlines the architectural framework and operational mechanics of the deep learning component, specifically the TF encoder. The Transformer architecture was initially conceived by Vaswani *et al*. [17] as a solution for natural language processing (NLP) challenges, distinctively characterized by its innovative integration of a self-attention mechanism. In recent periods, TF models have also been successfully adapted for forecasting wave height parameters [20]. In standard configurations, the network incorporates embedding layers to dimensionally expand word vector representations, alongside integrated encoder and decoder blocks designed to execute sequence-to-sequence translations. Nevertheless, because the present study constitutes a binary classification problem, only the isolated encoder apparatus is implemented.

As depicted in **Figure 7**, the left segment illustrates the schematic of the self-attention block, whereas the right segment displays the feed-forward neural network. The multi-head attention mechanism computes the interaction scores for the given inputs utilizing the following formulation:

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

Where Q, K, V are the matrices of queries, keys and values, the initial values of which are identical; and d_k is the dimension of the key matrix. As shown in **Eq. 1**, the multi-head attention layer creates several independent linear representations from the Q, K, V inputs. These representations are then subjected to layer normalization and merged with the original inputs via residual connections before entering the adjacent feed-forward structure. Additionally, a 1-D CNN is incorporated to augment the model's capacity to represent localized feature interdependencies.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

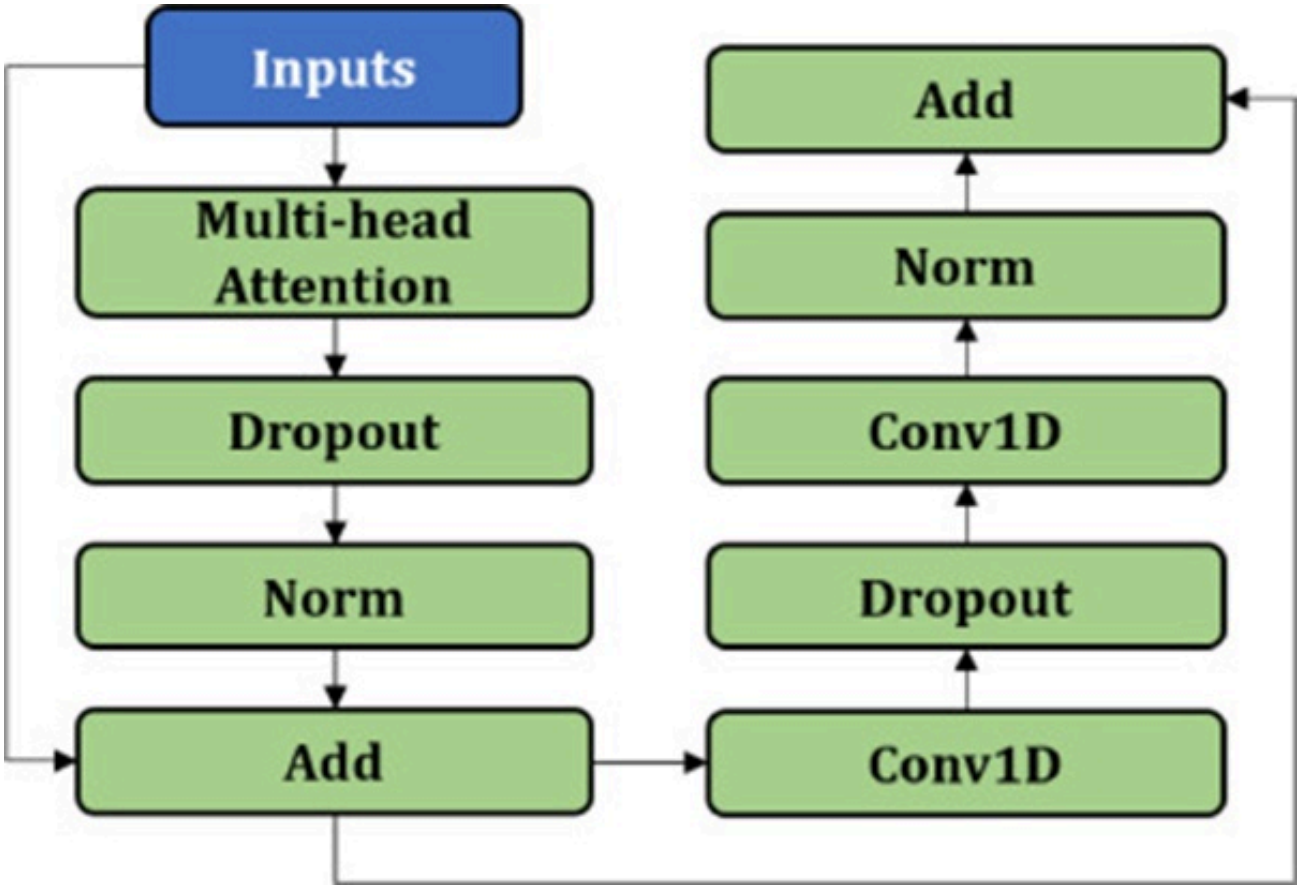


Figure 7. Structure of a single TF encoder used in the current study.

Training settings

This subsection delineates and elucidates the primary hyperparameters of the implemented model. Furthermore, it details the specific training protocols and the computational server configurations utilized.

As outlined in **Table 2**, both the head size d_k and the total number of heads h are established at 6 and 2, respectively. Sequential TF blocks are deployed in tandem to forge deeper inter-neuronal connections. The dimensionality of the initial filter within the 1D-CNN is determined by the product of d_k and h, featuring a highly compact kernel size configured to exactly one window per kernel.

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

Hyperparameter	Configured value
Attention head count	6
Dimension per head	6
Transformer block quantity	2
Number of 1D-CNN filters	256
1D-CNN kernel dimension	1

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

MLP unit count	128
Multi-head attention dropout rate	0.1
MLP dropout rate	0.4
Number of MLP layers	1
Batch size	32
Primary activation function	ReLU
Output layer activation function	softmax
Starting learning rate	0.001
Loss function	Sparse categorical cross-entropy

Optimization was executed using the Adam algorithm [24], initialized with a learning rate of 1e-3. To accelerate training and mitigate gradient vanishing during backpropagation, the rectified linear unit (ReLU) was selected as the activation function for all components excluding the output layer. A 2-element softmax function terminates the network architecture, yielding a probabilistic categorical output. Consequently, examining a singular data instance, the operational loss function is formulated as:

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

Where CCE denotes categorical cross-entropy; C (= 2) signifies the total count of classification categories; t_c and p_c represent the true observed state (integer) and the corresponding probabilistic prediction under classification category c, respectively.

The analytical operations and model training were conducted utilizing the Python 3.8 environment, leveraging the Keras framework [25]. All computational tasks were distributed across two cloud-based servers, each outfitted with an RTX 3070 GPU possessing 8 GB of video RAM.

Results and Discussion

This section presents the quantitative outcomes and multi-faceted analyses derived from comprehensive parallel comparisons. The comparative benchmark encompasses both EMD-integrated and non-EMD frameworks applied across mainstream architectures, specifically TCN, 1D-CNN, and LSTM.

Performance criteria

In conjunction with the ultimate average CCE (**Eq. 2**), sparse categorical accuracy (CA) serves as a primary metric for assessing the predictive efficacy of the proposed architectures. The mathematical expression for CA is defined in **Eq. 3:**

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

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

Where M_{test} designates the aggregate count of predicted labels (binary) that perfectly correspond with the true labels within the testing subset; and N_{test} represents the overarching dimensional size of the test dataset.

Accuracy comparison

This subsection explores the comparative accuracy across the models. The three parallel baseline models underwent systematic tuning, adhering to the fundamental parameters specified below:

1. TCN: Incorporates two discrete TCN layers utilizing 64 and 32 filters, respectively; Dilation rates are configured as [1, 4, 15, 32].

2. 1D-CNN: Comprises three sequential 1D convolutional layers, each equipped with 64 filters and a kernel size of 3, applying consistent ‘same’ padding at the boundaries.

3. LSTM: Utilizes two interconnected LSTM layers, each containing 32 units; the primary LSTM layer is configured to return the full sequence, whereas the secondary layer outputs strictly the final state.

Effects of lead time

Within this section, comparative accuracy evaluations are executed across varying forecasting lead times, spanning from 6 h to 48 h. To rigorously isolate the variables, a standard window size of 30 h is maintained across all trials, and a static sampling protocol is enforced to guarantee that the training and testing data subsets remain strictly invariant.

The performance trajectory for up to one day ahead is illustrated in **Figure 8** and **Table 3**. As the predictive lead time extends, the accuracy of all EMD-augmented models declines progressively. Notably, the performance curves for the TF and LSTM architectures are highly comparable, essentially converging. At the brief 6 h lead time, their respective accuracies on the test cohort are exceptionally high, achieving 0.991 and 0.988. They sustain robust performance levels approximating 0.900 even as the predictive horizon expands to a full 24 h, highlighting a pronounced capacity for assimilating distant temporal correlations. Conversely, models predicated on TCN and 1D-CNN experience a substantial degradation in precision as the lead time increases. While both initiate with accuracies exceeding 0.950, their ultimate precision deteriorates to approximately 0.800 at the 24 h forecast interval.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

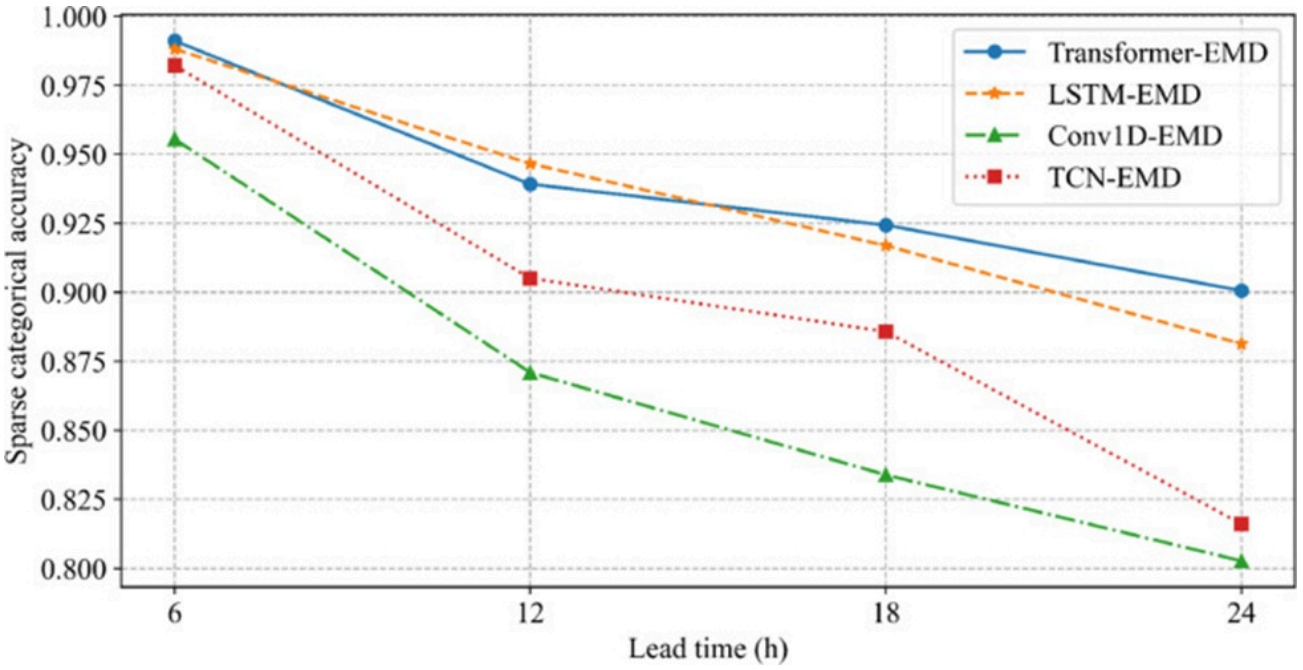


Figure 8. Model accuracy changes w.r.t. different lead time from 6 to 24 h.

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

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

Similarly, **Figure 9** and **Table 4** present the accuracy progressions extended to a two-day horizon, revealing a more pronounced precipitous decline. The absolute discrepancy between EMD-TF and EMD-LSTM remains minimal, with both concluding near a 0.720 accuracy mark at 48 h. Consistent with earlier observations, the TCN and 1D-CNN architectures underperform. However, the absolute accuracy differential between the TCN model and both the LSTM and TF models narrows. In contrast, the precision of the 1D-CNN model continues to deteriorate in direct response to the expanded lead time.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

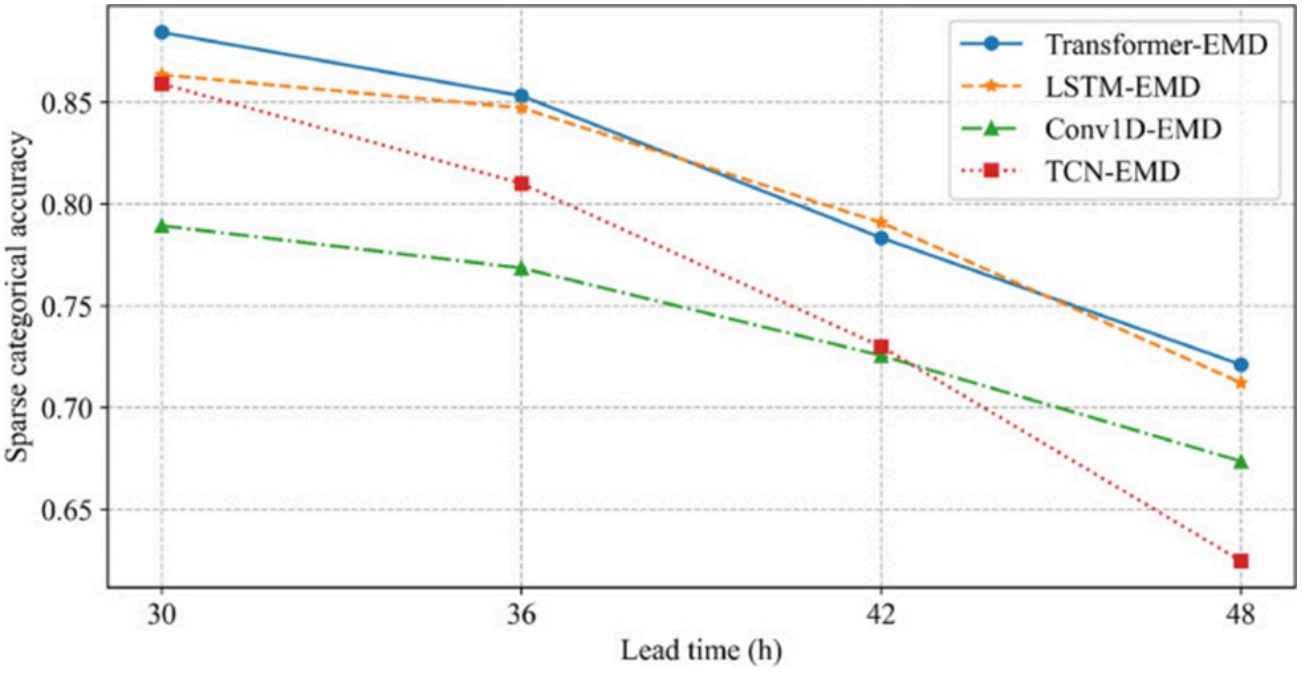


Figure 9. Model accuracy changes w.r.t. different lead times from 30 to 48 h.

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

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

The preceding evaluation confirms that across a continuous 2-day predictive window, the TF-based architecture secures an equivalent, if minimally superior, accuracy level compared to the LSTM framework. The predictive capability is profoundly strong during the initial 24 h but undergoes continuous attenuation thereafter. Concurrently, the exactitude of the newly proposed TF framework decisively eclipses that of the 1D-CNN and TCN configurations.

Effects of EMD features

This segment investigates the specific impact generated by the integration of EMD features. To magnify observational differences, models functioning at the absolute minimum (6 h) and maximum (48 h) lead times were isolated for this specific analysis.

As depicted in **Figure 10** and **Table 5**, under the 6 h forecasting scenario, EMD features exert a critical influence on minimizing model loss. The cross-entropy metrics across all evaluated models plummet precipitously from their baseline, non-EMD plateaus, which hover uniformly around 0.210. Notably, even the comparatively weakest architecture, the 1D-

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

CNN, registers a nearly 50% performance enhancement. However, the influence of EMD on the TF model is staggeringly prominent, achieving an almost 90% reduction in loss relative to its unmodified state.

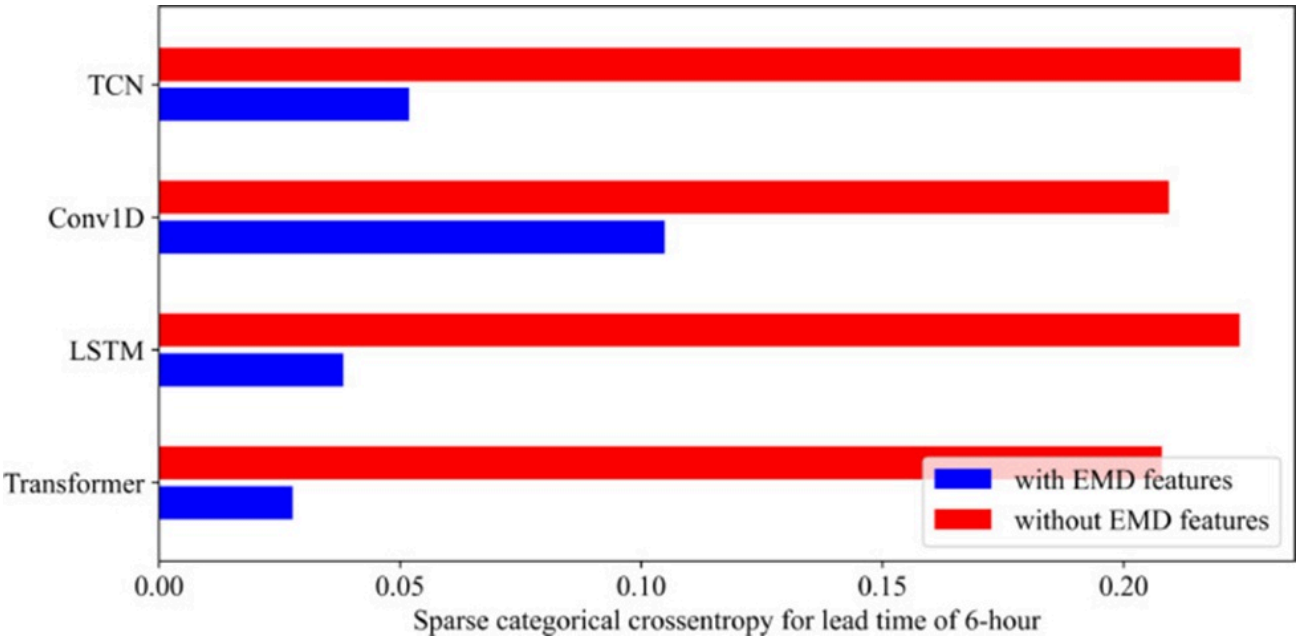


Figure 10. Loss chart comparisons for different models with or without EMD features under a lead time of 6 h.

Table 5. Loss value comparisons for different models with or without EMD features under a lead time of 6 h.

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

Conversely, analyzing **Figure 11** and **Table 6** regarding the 48 h predictive horizon, the advantageous impact of EMD features becomes markedly subdued, with the performance variance between EMD-enabled and non-EMD models restricted to roughly 25%. Importantly, within the non-EMD group, the isolated TF model does not yield the lowest absolute loss. Yet, upon integrating EMD features, the TF architecture shows a distinct, notable improvement over all competing models. This specific dynamic is likely attributable to the multi-head attention mechanisms inherent in the TF structure, which possess a superior capability to parse and leverage pertinent information embedded within the IMF sequences for extended future projections. Consequently, EMD features demonstrate a higher degree of operational synergy with the TF architecture than with alternative algorithms.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

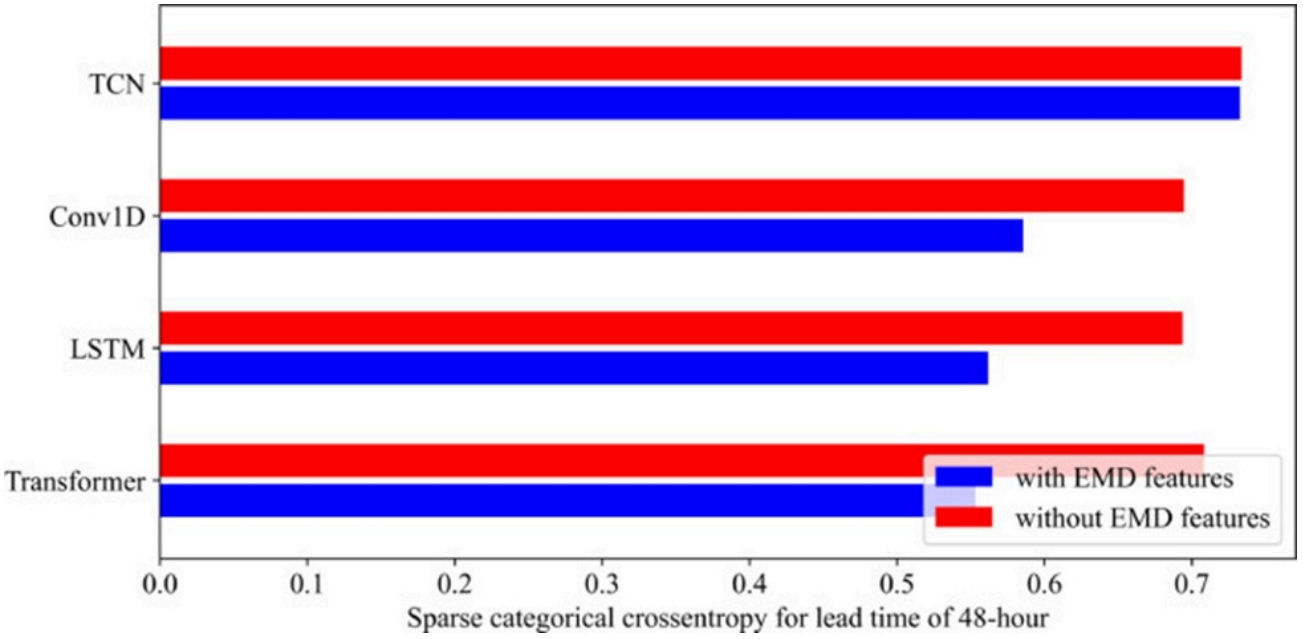


Figure 11. Loss chart comparisons for different models with or without EMD features under a lead time of 48 h.

Table 6. Loss value comparisons for different models with or without EMD features under a lead time of 48 h.

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

Effects of window size

Furthermore, the dimension of the input window—frequently referred to as the multi-step length in related literature—can significantly influence outcomes. Therefore, this subsection systematically evaluates computational loss across varying window sizes, stratified by distinct lead times.

As detailed in **Figure 12**, to scrutinize the models' generalization capabilities, multiple independent data samplings were processed while holding both lead time and window size constant (represented by the solid lines); the discrete dots delineate their corresponding mean loss values. Concurrently, the overlaid dashed lines trace the broader trajectory of cross-entropy fluctuation within the testing subsets as the window size progressively expands.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

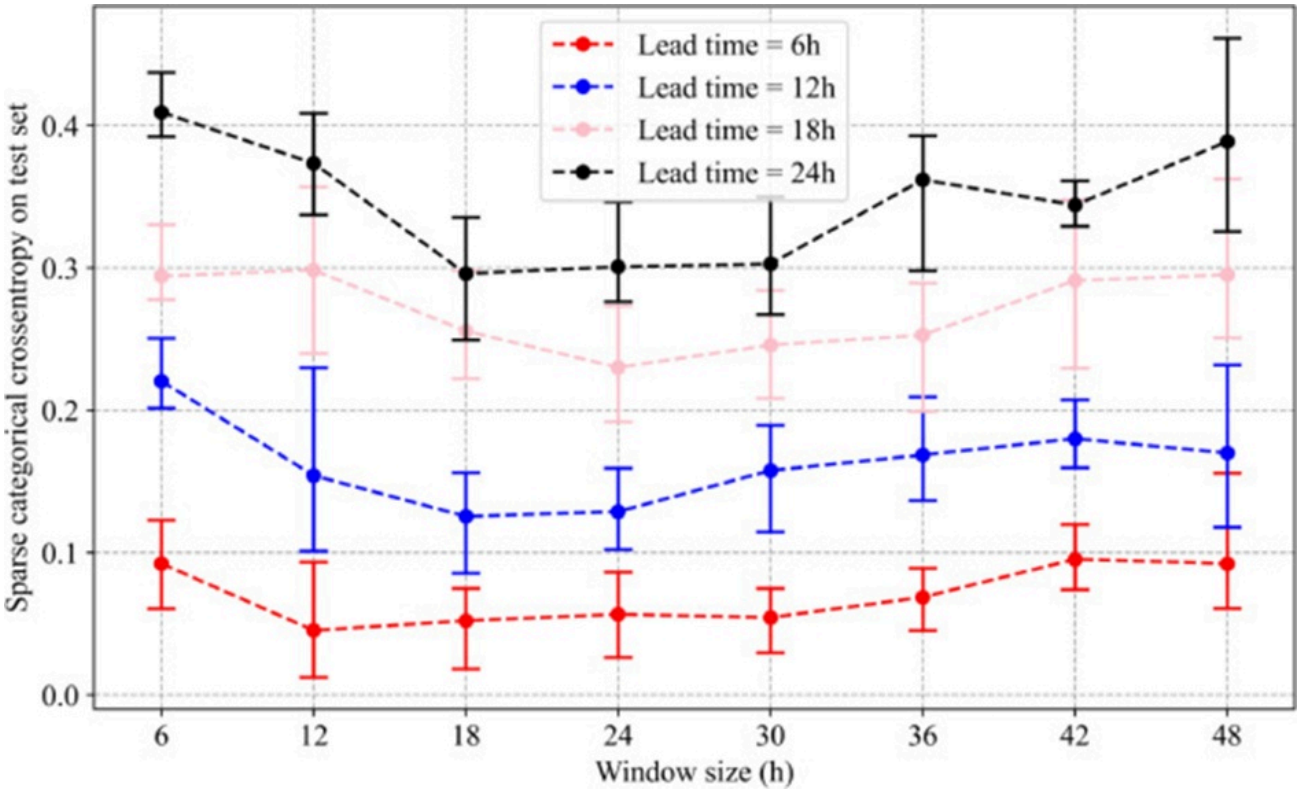


Figure 12. Test loss variation with different window sizes under the lead time changing from 6 h to 24 h.

Observations from **Figure 12** reveal that as the predictive lead time stretches from 6 h to 24 h, the magnitude of loss variation corresponding to distinct window sizes intensifies. Similarly, the dispersion range across disparate data samplings broadens in tandem with longer lead times. Crucially, the empirically optimal window size exhibits a gradual rightward shift: anchoring at 12 h for a 6 h lead time; adjusting to 18 h for a 12 h lead time; and stabilizing at 24 h for lead times of both 18 and 24 h. This phenomenon is logically predicated on the premise that extended forecasting horizons require a larger repository of historical data to accurately synthesize long-range dependencies. Conversely, for highly compressed lead times, temporally distant data points likely act as redundant noise, thereby exacerbating the training load and subsequently degrading overall accuracy.

Nevertheless, as the lead time is extended even further, these observed behavioral patterns begin to diverge, a shift documented in **Figure 13**. At a 30 h lead time, the previously established trends persist, characterized by a more pronounced amplitude of fluctuation across various window configurations. Furthermore, the theoretically optimal window size continues its rightward progression, aligning at 30 h as highlighted by the yellow markers. Strikingly, this progressive pattern abruptly terminates at the 36 h lead time threshold. Between the 36 h and 48 h marks, the loss trajectories uniformly and monotonically ascend in direct correlation with increasing window size, positioning the optimal window dimension strictly at the minimum evaluated step length of 6 h. A plausible hypothesis for this inversion is that most IMF features are inherently tuned to capture localized extreme behaviors. When the predictive horizon expands excessively, the localized developmental noise patterns reflected by these features lose their correlative significance. Consequently, as the input window dilates, this influx of non-contributory data merely complicates the training process, actively fostering an environment highly susceptible to overfitting.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

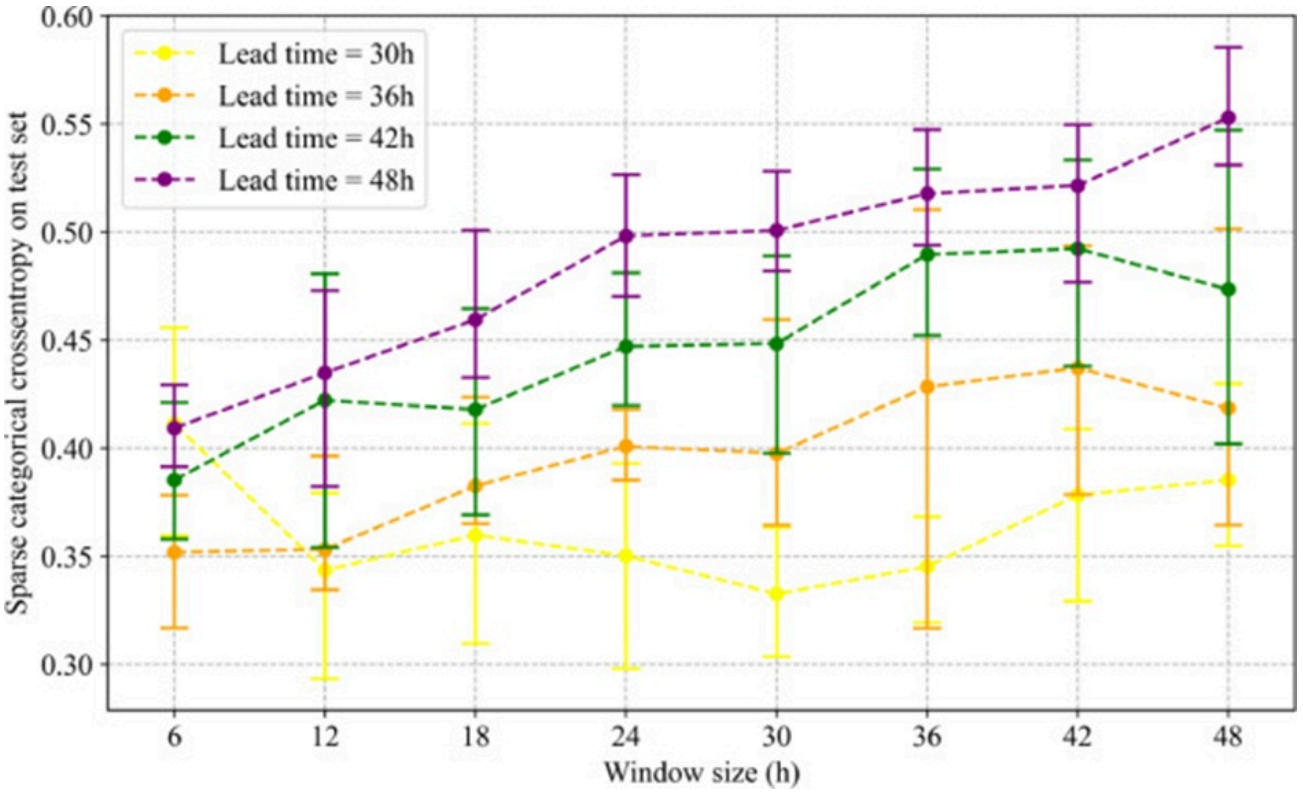


Figure 13. Test loss variation with different window sizes under the lead time changing from 30 h to 48 h.

Efficiency comparison

Beyond accuracy, efficiency plays a crucial role in model evaluation, particularly for real-time dynamic applications. In this study, representative cases of 30 (lead time) / 30 (window size) are utilized to assess the computational time required by various EMD models. Because of variations in the sampling datasets, the average overall computing time and the number of epochs needed to achieve the optimal model utilizing the early-stopping technique are presented in **Table 7.**

Table 7. Averaged total time cost and epoch number comparison between different EMD models.

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

As demonstrated in **Table 7**, the overall computational time of the TF-based model places it second among the four models examined, requiring roughly twice the time of the fastest model, EMD-Conv1D. The disadvantages of the EMD-LSTM are clear. Firstly, although all four models converge within an acceptable number of steps, the EMD-LSTM still demands a relatively higher number of epochs. Secondly, the total time expenditure of the EMD-LSTM significantly exceeds that of all other models, taking more than ten times longer than the EMD-TF due to its inherently serial

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

architecture. Consequently, when factoring in the accuracy results, the EMD-TF emerges as a highly efficient and promising approach for this specific task.

To implement the proposed model in practical scenarios, establishing collaborative online and offline systems is essential. The online component will gather historical sensor measurements, utilize the deployed model to perform classifications, and persistently relay new data to the offline component. Concurrently, the offline system will routinely retrain the model according to predetermined error thresholds and return the refined model to the online system for active deployment. Through this synchronized framework, the model can undergo continuous optimization and improvement, yielding progressively more precise and dependable predictions across various SWH classification operations.

Conclusion

This research introduces a novel framework known as empirical mode decomposition with Transformer encoder (EMD-TF), designed to enhance wave energy management through the advanced classification of significant wave height levels. For this investigation, the upper and lower thresholds for significant wave height (SWH) are established at 2.0 m and 1.5 m, respectively. Based on numerical experiments, the proposed model demonstrates the following notable benefits:

1. Within lead times of 24 h and 48 h, the proposed model achieves an accuracy up to 9.80% and 9.97% greater than that of temporal convolutional network and one-dimensional convolutional neural network-based approaches. The performance of the long short-term memory network combined with empirical mode decomposition (EMD-LSTM) is comparable, though it remains slightly less accurate than the proposed model.

2. Given a brief lead time of 6 h, integrating features derived from empirical mode decomposition significantly decreases the testing loss across all parallel models. However, this approach proves most impactful when applied to Transformer models, resulting in a nearly 90% reduction in categorical cross-entropy.

3. When evaluating extended lead times up to 48 h, the impact of features extracted via empirical mode decomposition is considerably diminished, yet it remains most noticeable within the Transformer architecture.

4. The proposed model exhibits distinct behavioral patterns concerning the impact of window size across varying lead time intervals. For lead times shorter than 30 h, the optimal window size tends to shift rightward, whereas the smallest available window size proves most effective for lead times exceeding 30 h.

5. Evaluations of computational efficiency reveal that the EMD-LSTM requires more than ten times the temporal cost of the EMD-TF model, indicating that the proposed framework serves as a highly efficient alternative for SWH classification operations.

It must be noted that the data utilized in this study are limited to the North Atlantic Ocean. Nevertheless, it is expected that the introduced methodology possesses the adaptability required for application to diverse datasets across global regions. Such adaptation can be accomplished by adjusting the upper and lower SWH thresholds to align with specific local environmental characteristics. Although the relationships between model performance and lead time might differ across various geographical areas, the underlying trends and overall efficacy of the framework are anticipated to remain consistent.

Moving forward, additional studies are necessary to comprehensively comprehend frequency or phase [26] attributes and their corresponding neural network mechanisms, including weight distribution, thereby enabling the development of more solid theoretical explanations.

Nomenclature

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

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

loT = 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

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

N_{test} = Total number of samples in test set

Acknowledgements

None

Conflict of interest

None

Financial support

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

Ethics statement

None

Received: 15 Jul 2023 Revised: 20 Sep 2023 Accepted: 06 Oct 2023
Published online: 15 June 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

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. Gómez-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. Abdullah 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; 2022:3828-32.

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, et al. 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, et al. 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 (PM2.5). 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, et al. 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, et al. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. Energy Convers Manage. 2021;227:113559.

17. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. 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, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond Ser A Math Phys Eng Sci. 1998;454:903-95.

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; 2022:430-4.

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. 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.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press

24. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.

25. Gulli A, Pal S. Deep Learning with Keras. Packt Publishing Ltd; 2017.

26. Xu X, Hu S, Shi P, Shao H, Li R, Li Z, et al. 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.

© 2025 The Author(s). | Published by Institute of Smart Energy Systems Press