Ocean Engineering 242 (2021) 110130

Contents lists available at ScienceDirect

Ocean Engineering

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

Nowcasting significant wave height by hierarchical machine learning classification

Demetris Demetriou [a,*], Constantine Michailides [a], George Papanastasiou [b], Toula Onoufriou [a]

[a] Department of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol, Cyprus
[b] VTT Vasiliko Ltd., Larnaca, Cyprus

ARTICLE INFO

Keywords:
Hierarchical machine learning
Classification algorithms
Significant wave height prediction
Classification based modeling
Hierarchical decomposition
Ocean engineering

ABSTRACT

This paper proposes an alternative method for nowcasting significant wave height (H_s) through the development of hierarchical machine learning classification models. In testing the hypothesis that hierarchical classification can improve H_s prediction, flat and hierarchical classifiers were developed and tested on field-data recorded on a coastal jetty located in the southern coasts of Cyprus. A comprehensive investigation of the performance of flat over hierarchical classification models yields that the proposed method provides greater flexibility throughout the model development stages. This flexibility is attributed to the manipulation of data before training, optimization of classifier's hyperparameters during training, and the curtailment of features post-training at each level of the hierarchy. It is demonstrated that, the hierarchical approach resulted in better classification performance across a plethora of performance metrics established for a comprehensive comparison. It is also shown that the increased performance of the proposed approach comes at the expense of complexity arising from performing computationally expensive operations and the requirement for development of multiple local classifiers. Still, the increased classification performance of the hierarchical approach highlights the potential of this original method and the requirement for a rigid framework to be constructed for the development of hierarchical models for H_s prediction.

1. Introduction

Nowcasting of wave refers to the prediction of wave parameters, typically height and period, in real-time. The terms significant wave height (H_s) and significant wave period (T_s), which are defined as the average of the largest one-third of the observed wave values, are used in many practical applications within coastal and offshore engineering. Owing to the importance of nowcasting these critical parameters, many assets are found to be equipped with wave measuring devices such as wave buoys and wave radars (Bolton, 2014). Between the two types of wave monitoring devices, a buoy has the advantage of being placed in open waters away from a structure, nullifying the wave-structure interaction that may affect the measured values. However, because of the nature of the device, a buoy can be easily damaged or even stolen. On the contrary, a wave radar which necessitates physical coupling to the structure, makes the readings prone to errors as a result of wave-structure interaction. Still, regardless the choice of device, there exists a big financial burden to be considered when acquiring, installing and maintaining these devices.

At the absence of sophisticated and expensive wave recording devices, nowcasting of waves is performed on the basis of numerical models such as the SWAN (Ou et al., 2002; Rogers et al., 2007), WAM (Group, 1988; Hersbach and Janssen, 1999; Monbaliu et al., 2000), Wave Watch III (Mentaschi et al., 2015) and JONSWAP (Hasselmann et al., 1973). Despite the complexity arising from solving the mathematically challenging physical equations that inevitably raise the computational burden, these methods have been deemed by many researchers not to be of the desirable accuracy (Browne et al., 2007; Zijlema, 2010), and do not generalize over various oceans and seas (Kumar et al., 2018) since they require accurate local bathymetric surveys, making their implementation a non-trivial task (Browne et al., 2007). As a result, the spatial resolution and consequent accuracy in the representation of coastal/ocean features in physics based models plays a major role in the overall accuracy and computational demands of wave simulations.

Owing to the shortcoming of traditional numerical methods, over the years, data driven models based on machine learning methods have received great attention by the research community as they are

* Corresponding author.
E-mail address: demetris.demetriou@cut.ac.cy (D. Demetriou).

https://doi.org/10.1016/j.oceaneng.2021.110130
Received 10 August 2021; Received in revised form 22 October 2021; Accepted 29 October 2021
Available online 5 November 2021
0029-8018/© 2021 Elsevier Ltd. All rights reserved.

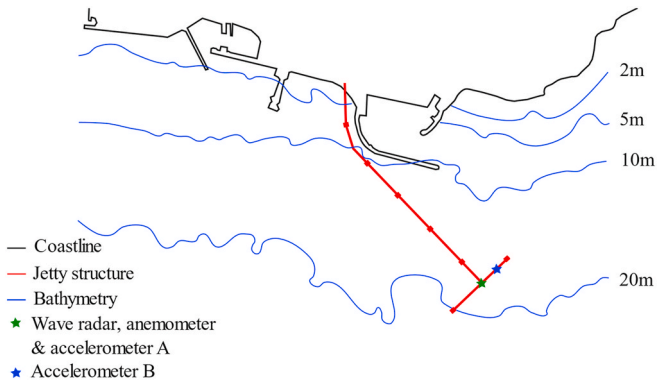

Fig. 1. Seabed bathymetry at the Vassiliko bay area and location of the installed monitoring equipment on the jetty.

Table 1
Summary of features extracted from wind and structural acceleration data.

Algebraic mean (MEAN)	Root mean square (RMS)	Standard deviation (STD)	Absolute signal energy (ASE)
Wind Speed	Acceleration Ax	Acceleration Ax	Acceleration Ax
Wind Speed $_{-1}$	Acceleration Ay	Acceleration Ay	Acceleration Ay
Wind Speed $_{-3}$	Acceleration Az	Acceleration Az	Acceleration Az
Wind Speed $_{-6}$			
	Acceleration Bx	Acceleration Bx	Acceleration Bx
Wind Direction	Acceleration By	Acceleration By	Acceleration By
Wind Direction $_{-1}$	Acceleration Bz	Acceleration Bz	Acceleration Bz
Wind Direction $_{-3}$			
Wind Direction $_{-6}$		Wind Speed	
		Wind Direction	

generally considered a great alternative to explicit numerical models because of their computational efficiency and reported accuracy. Machine learning wave models scale down the wave prediction problem from a regional to a local scale, and contrary to explicit numerical wave models they do not require any spatiotemporal information in their development, as they are based on mapping the input-output relationship of data. In this respect, exploiting the input output relationships of wind and wave data, Neural Network (NN) (Browne et al., 2007; Deo et al., 2001; Gunaydın, 2008; Rao and Mandal, 2005; Tolman et al., 2005), Decision Tree (Mahjoobi et al., 2008), Random Forest (Sharma and Jain, 2018), Support Vector Machine (Mahjoobi and Mosabbeb, 2009), Fuzzy Inference (Wu et al., 2019), Bayesian Interference (Egozcue et al., 2005; Scotto and Soares, 2007), and K-nearest neighbor (KNN) (Wei, 2018) models have been developed for the prediction of significant wave height with fluctuating accuracy depending on the choice of algorithm. A comparison of the performance of different machine learning algorithms on the prediction of H_s can be found in the recent studies of (LI and Liu, 2020; Shamshirband et al., 2020; Wang et al., 2021).

Due to the nature of the H_s prediction problem, the studies presented above involved the use of machine learning algorithms in a regression predictive framework, based on approximating a mapping function from input variables to a continuous output variable space. Most recently, a new era of H_s prediction models based on classification predictive modeling (CPM) were proposed by the authors of the present manuscript. Contrary to regression, CPM involves the input/output mapping to discrete output variables, often termed classes or labels. In their study, Demetriou et al. (2021) found that classification models trained on wind speed and direction, as well as structural acceleration data in two points, recorded on a coastal jetty structure, can yield models of enhanced H_s predictive capacity, suggesting that a) structural data can drastically influence the performance of the wave predictive models and b) classification algorithms can be successfully applied in the nowcasting of H_s. The latter observation advocates that a new area of research can be exploited, and tools developed for classification learning in other areas of the machine learning domain (e.g. computer vision, text categorization, music genre classification etc.) could be used to improve H_s models.

In this respect, the present study aims to exploit the concept of CPM of H_s by proposing an extension of the conventional flat classification problem approach presented in the study of Demetriou et al. (2021) to one that embraces the hierarchical development of the machine learning classification network for increasing the flexibility in the overall system. The initial hypothesis entails that transitioning from flat to hierarchical classification can aid with; a) reducing the bias of predictive models owing to imbalanced datasets, b) improving algorithm flexibility by allowing optimizing algorithm hyperparameters and selecting optimal feature subset depending on the wave height range in the hierarchy, and c) reducing computational demands attributed to the lower feature dimensionality of the models during training. The ultimate objective of this study is to prescribe a transferable methodology that allows the development of optimized classification models with enhanced performance both in terms of predictive accuracy and computational efficiency for the area of significant wave height prediction.

2. Study area and instrumentation

For investigating the performance of wave models developed via the flat and hierarchical classification approach, the monitoring equipment installed on a fixed-bottom coastal jetty structure located in the Southern coasts of Cyprus is used to provide the data to be used in the training and testing of supervised machine learning models. To this end, data collection was performed through the following equipment, namely, an anemometer, two triaxial accelerometers and a wave radar. The accuracy sensitivity of the accelerometers and the wave radar are ±0.4 dB max over the bandwidth and ±6 mm respectively. It is noteworthy that all the equipment installed on the jetty is operated by VTT Vassiliko Ltd. for operational purposes.

The anemometer's speed and direction were recorded at 14 m above sea level, while the wave radar was located at the T-junction part of the jetty, 10 m above sea level. The water depth at the location of the wave radar is 18.70 m. It is noted that the seabed topography of the Vassiliko bay area is relatively uniform with smooth gradients and no significant irregularities or obstructions. The seabed bathymetry at the Vassiliko bay area and the location of the monitoring equipment installed on the jetty is exhibited in Fig. 1. It is noted that bathymetry data in the exhibited figure are extracted from the study of Bergin and Spearman (2006).

Acceleration data were acquired from two triaxial accelerometers installed on the coastal jetty. One accelerometer is located at the T-junction of jetty and a second accelerometer is installed on the bridging elements of the T-junction and the main structural system.

Wave data were sampled at single minute intervals (0.0167 Hz), while acceleration and wind data were recorded at a sampling frequency of 8 Hz. The sampling frequency of 8 Hz for the case of the accelerometers was selected after numerically calculating the first two dominant structural modes of the T-junction of the jetty, found to be at 1.03 Hz and 2.52 Hz in the X-direction and 1.02 Hz and 3.36 Hz in the Y-direction respectively (Demetriou et al., 2021). Accordingly, a sampling frequency of 8 Hz was selected for satisfying the Nyquist criterion, which states that a waveform can be correctly reconstructed provided that the sampling frequency of the instruments is greater than double the highest frequency to be sampled.

For the compatibility of the acceleration data with the sampling frequencies of the wind and wave data, data condensation was performed on the wind and acceleration dataset in single minute windows. To this end, at each minute window, the recorded wind and acceleration values were condensed by extracting their algebraic mean (MEAN), root

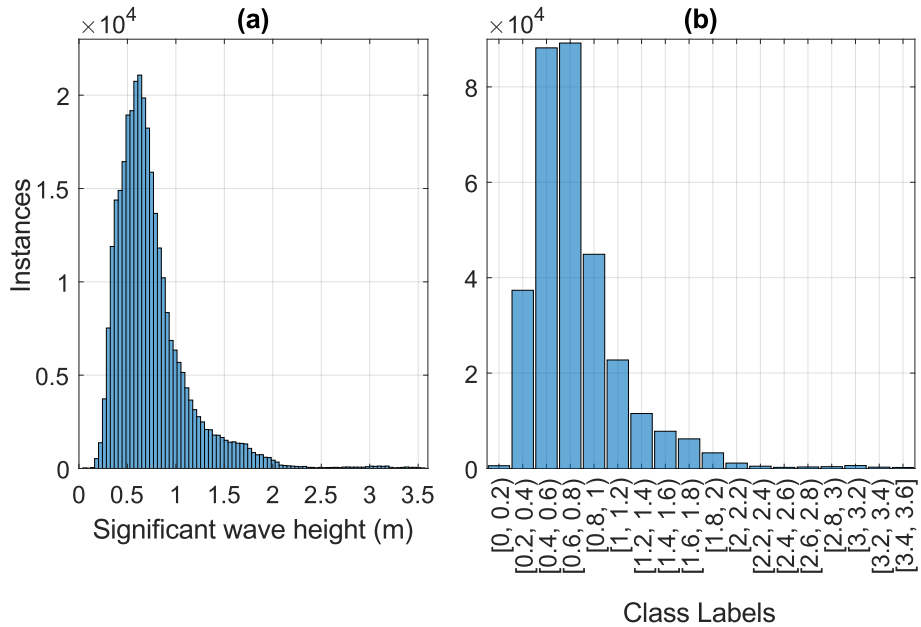

Fig. 2. (a) Distribution of significant wave height at the recording location, (b) distribution of class labels in the flat classification approach.

mean square (RMS), standard deviation (STD) and absolute signal energy (ASE). The resulting dataset comprised of over 320,000 single minute samples of wind speed, wind direction, structural acceleration and significant wave height data recorded over the period of 2020–2021. Table 1 summarizes the features used in the training of the machine learning models. Accordingly, the input matrix attained the size of N × 28 (N ≈ 320,000) with 10 features being attributed to wind data and 18 features being attributed to structural data. Similarly, the output matrix attained the size of N × 1 containing N samples (N ≈ 320,000) of the recorded significant wave height values. With reference to Tables 1, 8 wind derived features correspond to the MEAN values of wind speed and direction at the recording time of the output variable, as well as 1, 3 and 6 h a-priori. This is noted by the subscripts used in Table 1 (e.g. Wind Speed$_{-1}$ etc.). Similarly, the remaining two wind derived features correspond to the STD of wind speed and direction at the recording time of the output variable. The remaining 18 features correspond to the RMS, STD, and ASE for the two accelerometers in each of the three orthogonal directions at the recording time of the output variable. Adopting the notation presented in Table 1, Acceleration Ax corresponds to the acceleration values recorded by the first accelerometer (e.g. the accelerometer located on the jetty's T-junction) in the X-direction. Similarly, Acceleration By corresponds to the acceleration values recorded by the second accelerometer (e.g. the accelerometer located on the jetty's bridging elements) in the Y-direction and Acceleration Bz corresponds to the acceleration values recorded by the second accelerometer in the Z-direction.

2.1. Imbalanced significant wave height datasets

An imbalanced dataset occurs when the distribution of samples across the known classes is skewed or biased. Indeed, most real-world applications are often characterized by imbalanced datasets, where the well-represented class is called the majority and the under-represented class is called the minority class. Depending on the ratio of samples in each of the classes, the imbalance can be slight or severe.

This skewness in the dataset can degrade the predictive capacity of trained models, as most machine learning classification algorithms operate under the assumption of equally distributed classes with the resulting classifiers being typically biased towards the majority class (Demetriou et al., 2021; García and Herrera, 2009; Tang et al., 2009). The performance of classification models deteriorates further when imbalanced datasets contain overlapping samples in the feature space, making the classification of each class indistinguishable (Das et al., 2014). It has been also observed that imbalanced datasets with the minority class containing low sample numbers and high-dimensionality features can cause further deterioration in the predictive performance of the classifier (Japkowicz and Stephen, 2002). This performance deterioration is amplified when large imbalanced datasets are considered (Buda et al., 2018; Takerkhani et al., 2020).

Literature prescribed measures to account for data skewness in classifier training include cost-sensitive learning (Zheng and Zhao, 2020), and selection of algorithms that are capable of coping with such uneven distributions. The Synthetic Minority Oversampling Technique (SMOTE) and the Random Under-Sampling (RUS) method implemented via the SMOTEBoost (Chawla et al., 2003) and RUSBoost (Seiffert et al., 2008) algorithms, constitute great examples of the approach of resampling the data by under-sampling the majority and over-sampling the minority class by cropping and creating new synthetic data respectively.

Evidently, cost-sensitive learning becomes increasingly difficult to implement properly and is highly dependent on the characteristics of the dataset (Beyan and Fisher, 2015). For this reason, cost-sensitive learning becomes particularly complex and extremely convoluted for the case of multi-class classification problems such as the one presented herein. Complementing the aforementioned, it has been also reported that cost-sensitive learning along with resampling techniques (SMOTE or similar) do not always relay to classification performance gains over the original classifiers (Han et al., 2005), potentially owing to the newly created synthetic samples causing overlapping of the minority and the majority classes.

It is therefore important to highlight that for the development of a successful classifier capable of dealing with the case of imbalanced datasets, appropriate data manipulation and feature selection needs to occur prior to model training, while post classification, appropriate evaluation metrics need to be established to account for the bias towards the majority class. Demetriou et al. (2021) discussed extensively the issue arising from the use of the Accuracy metric (the sum of correctly predicted minority and majority samples over the total samples) in the classification of significant wave heights and highlighted the necessity of using alternative metrics including Markedness and Informedness (Powers, 2007).

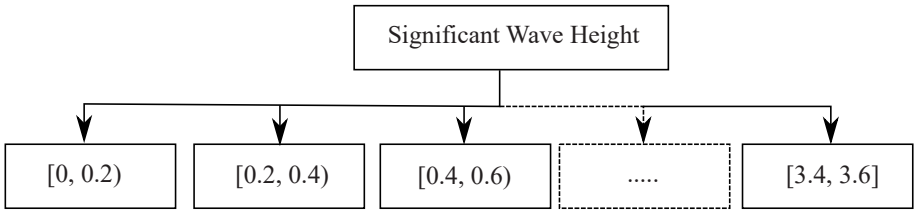

Fig. 3. Flat classification of significant wave height in accordance to the study of Demetriou et al. (2021).

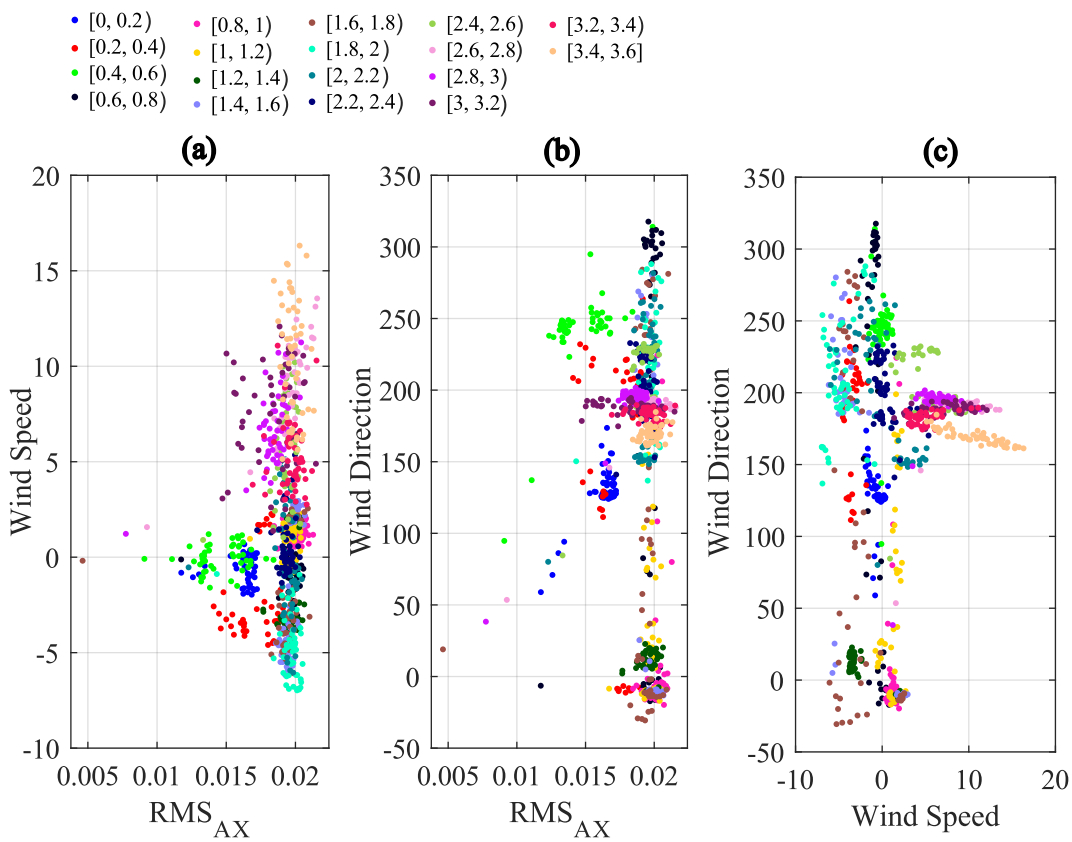

Fig. 4. Scatter plots of (a) Wind Speed over RMS_{AX}, (b) Wind direction over RMS_{AX}, and (c) Wind direction over Wind Speed, for the case of flat classification.

To exhibit the imbalanced nature of the significant wave height prediction problem, Fig. 2(a) depicts the distribution of the recorded significant wave height at the Vassiliko bay area, while Fig. 2(b) exhibits the same distribution categorized in 18 bins of 0.2 m in accordance to Demetriou et al. (2021), representing each of the 18 classes used as labels in the supervised machine learning classification. It is evident from Fig. 2(a) and (b) that the particular dataset is highly imbalanced, and a classifier will typically tend to be biased towards the majority class (e.g. [0.6, 0.8)). Additionally, it is evident that the specific problem falls within the category of multiclass classification as the developed classifier will have to distinguish between a total of 18 classes, making this particular multiclass classification problem of imbalanced datasets rather ambitious.

In the following sections, this paper presents a systematic approach for dealing with the issue of imbalanced datasets for the case of significant wave height prediction. The proposed approach falls outside the complex area of cost-sensitive learning and data resampling techniques. Rather, the proposed approach exploits the merits of classification learning, and more specifically the advantages of using hierarchical decomposition of significant wave heights.

2.2. *Flat vs hierarchical significant wave height classification*

In order to understand the notion of hierarchical classification of H_s, it is important to firstly define its flat classification counterpart. Flat classification involves no inherent hierarchy between the possible categories that data can belong to. In the problem of H_s classification, the objective is to classify H_s in a number of distinct and predetermined classes. In the study of Demetriou et al. (2021) the authors arbitrarily decomposed the classification problem into 18 distinct classes of 0.2 m bins ranging from 0 to 3.6 m (Fig. 2(b)). The adopted flat classification approach is schematically presented in Fig. 3.

This approach has the serious disadvantage of having to build a single classifier to discriminate amongst many classes, without exploring information about parent-child class relationships present in the class hierarchy (Silla and Freitas, 2011). Importantly, this sole classifier needs to address the classification problem without being biased towards the most represented class attributed to the highly imbalanced nature of the dataset. As discussed in the preceding sections of the manuscript, particularly for the cases of large, imbalanced datasets, and high dimensionality features the performance of the flat classifier deteriorates significantly if class overlap is present. Fig. 4 exhibits the relationship between 3 out of 28 selected features used to train the flat

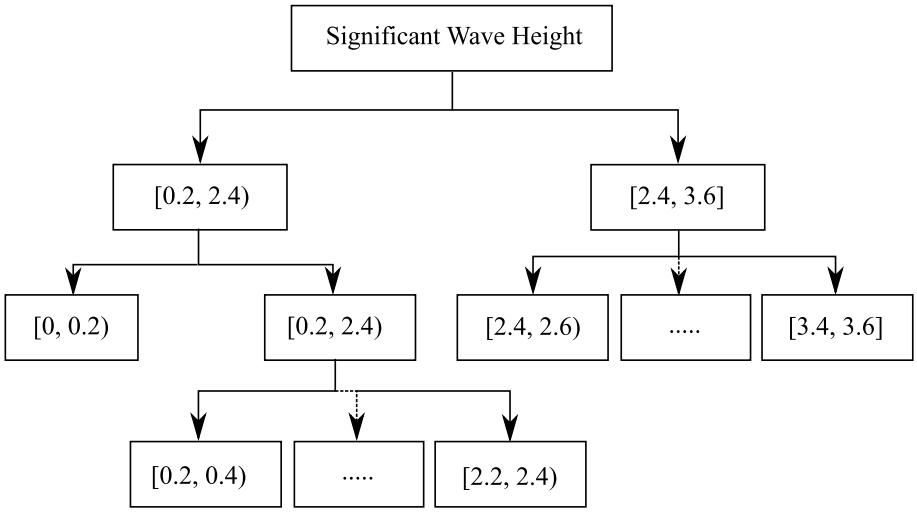

Fig. 5. Hierarchical classification of significant wave height.

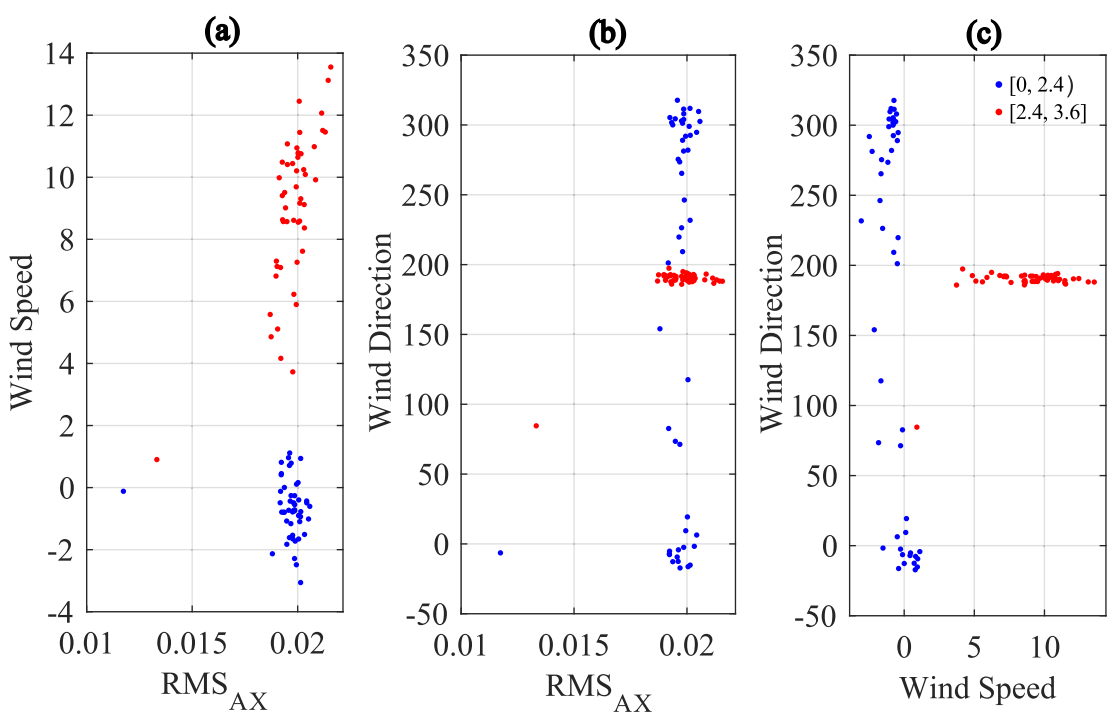

Fig. 6. Scatter plots of (a) Wind Speed over RMS_{AX}, (b) Wind direction over RMS_{AX}, and (c) Wind direction over Wind Speed, at the first (higher) level of the hierarchical structure.

classifier in Demetriou et al. (2021), namely, MEAN wind speed, MEAN wind direction and RMS acceleration of the first accelerometer in the X-direction (denoted RMS_{AX} in Fig. 4 and subsequent figures) presenting 100 random samples attributed to each of the 18 classes. Fig. 4 exhibits an apparent class overlap present in the feature space, suggesting that extracting insights into feature importance is arguably impractical through the framework of flat classification. Additionally, owing to class overlap, balancing the imbalanced dataset by creating synthetic data through SMOTE or similar techniques will not aid classification.

On the contrary, there are several reasons why hierarchical decomposition might be beneficial in significant wave height classification. An example of a class hierarchy breakdown for the problem of significant wave height classification is demonstrated in Fig. 5. With reference to Fig. 5, significant wave heights are decomposed in a broader range of values at the higher levels of the hierarchy, narrowing down the band of significant wave height values at the lower levels of the hierarchy. The final level of the hierarchical structure consists of only leaf - classes (e.g. the classes would not be subclasses further), arranged in 0.2 m bins, similar to the case of flat classification.

It can be understood that hierarchical decomposition naturally reduces the ratio between the minority and majority class at each level of the hierarchy and as a result the reduced imbalance in the dataset can yield less biased classification models. At the same time, a hierarchical approach is favorable in terms of dividing the problem into manageable portions through separating the feature-space of each group into lower dimensions. This naturally reduces both training times and prediction speeds. Furthermore, hierarchical break-down of the classes allows model developers to extract insights into feature importance aiding with the removal of feature redundancy in model training. It is observed by many researchers that different features are better at discriminating

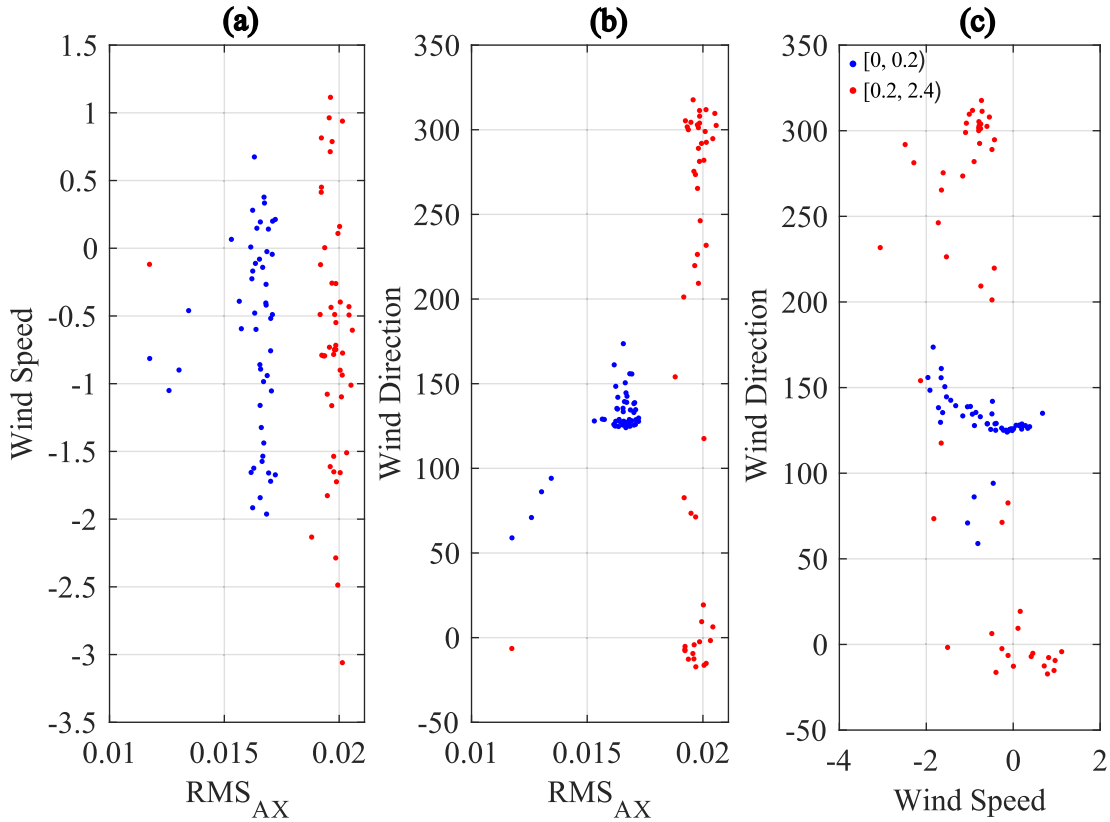

Fig. 7. Scatter plots of (a) Wind Speed over RMS_{AX}, (b) Wind direction over RMS_{AX}, and (c) Wind direction over Wind Speed, at the second level of the hierarchical structure.

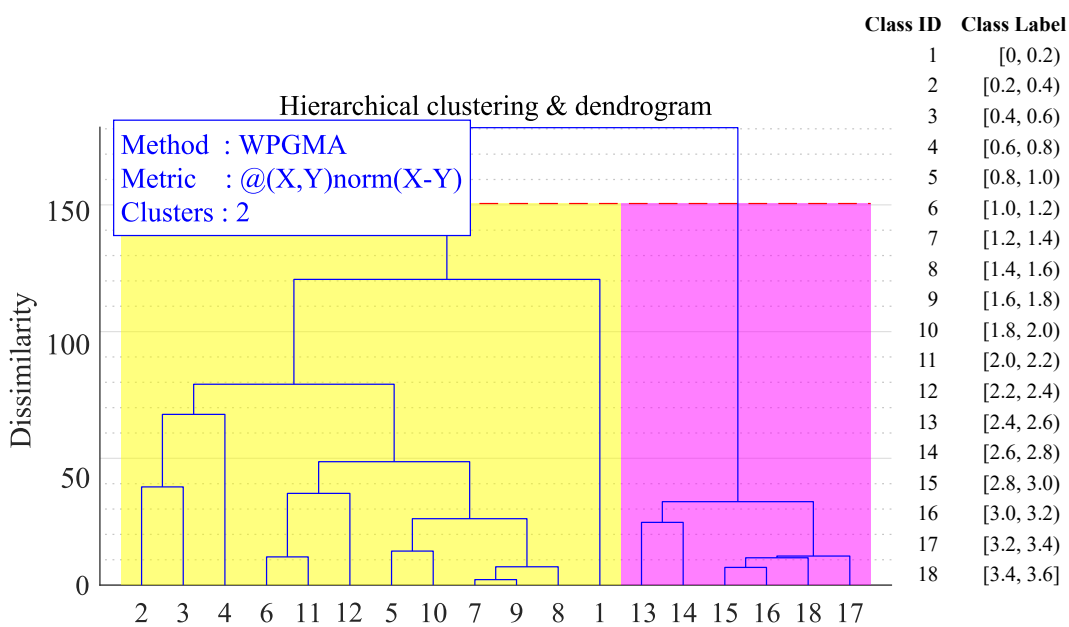

Fig. 8. Unsupervised clustering of significant wave height.

between classes at different levels of the class hierarchy (Ceci and Malerba, 2007; Koller and Sahami, 1997; Malakar et al., 2020; Secker et al., 2010). This is the result of selecting more specific domain features to a sub-problem which can increase the accuracy and efficiency of the classification models (Hao et al., 2007). For the case of H_s, this is exemplified by the influence of certain features at different wave height ranges (for example structural acceleration related features being more influential at large wave height ranges). In this regard, it can be hypothesized that omitting structural acceleration features at lower wave height ranges could yield less biased models. What is more, contrary to flat classifiers, hierarchical classification models provide insights of the performance of each local classifier. That is, the hierarchical approach

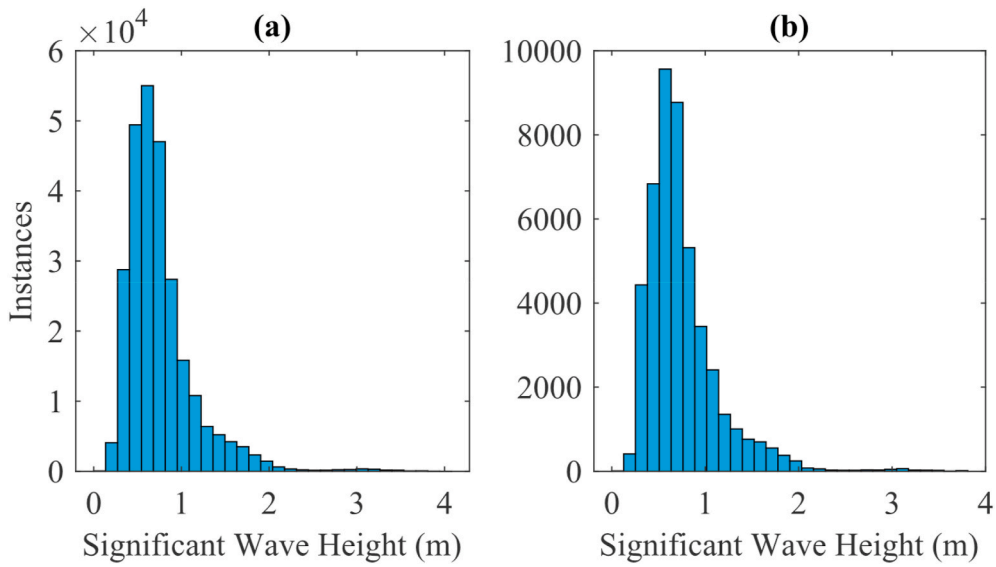

Fig. 9. Distribution of samples in (a) the training dataset and (b) testing dataset.

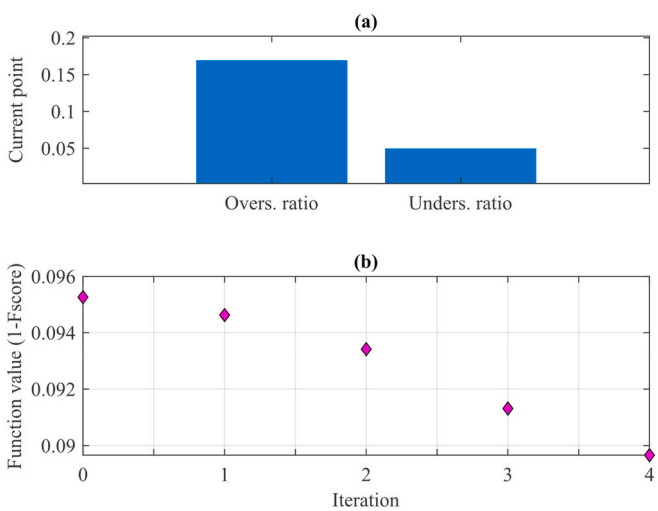

Fig. 10. (a) Optimized oversampling and under-sampling ratios, (b) achieved function output at each iteration from a single solver.

provides accuracy metrics for sets that share a common ancestor along the class hierarchy. This information may be crucial for model developers that may be less interested in the overall accuracy of a model and more by its ability to provide reliable information on specific ranges of the problem domain (e.g., how well the model classifies large significant waves). Finally, traditional class balancing techniques based on cost-sensitive analysis and resampling are easier to implement at each level of the hierarchy owing to the reduced problem dimensionality as well as the reduction in class overlapping. Fig. 6 exhibits the first level of classification in the hierarchical classification problem presented in Fig. 5. It is observed, that at the first level of this arbitrary hierarchical structure, class overlap is minimized compared to the feature class distribution observed in the flat approach exhibited in Fig. 4. Similar remarks can be made by investigating class distribution at the second level

Table 3
Summary of classifier training times post-optimization.

	Classifier	Algorithm	Training Data Size	Training Time Required
Hierarchical classifiers	Binary 1	GentleBoost	53 MB	39.44 s
	Binary 2	GentleBoost	54 MB	159.36 s
	Binary 3	LogitBoost	54 MB	246.60 s
	Multiclass 1	AdaBoost	28 MB	615.17 s
	Multiclass 2	AdaBoost	15 MB	239.14 s
	Multiclass 3	AdaBoost	335 KB	6.19 s
Flat Classifier	Flat classifier	AdaBoost	65 MB	2138.9 s

Table 2
Summary of classifier hyperparameters post-optimization.

	Classifier	Algorithm	No. Of Splits	No. Of Learners	Learning Rate	SMOTE
Hierarchical classifiers	Binary 1	GentleBoost	113	30	0.1	Yes Overs.Ratio:5% Undes.Ratio:0.1%
	Binary 2	GentleBoost	54	130	0.358	Yes Overs.Ratio:343% Undes.Ratio:25%
	Binary 3	LogitBoost	13,814	90	0.99	Yes Overs.Ratio: 17% Undes.Ratio: 5%
	Multiclass 1	AdaBoost	4900	330	0.95	No
	Multiclass 2	AdaBoost	9009	150	0.99	No
	Multiclass 3	AdaBoost	178	131	0.99	No
Flat Classifier	Flat classifier	AdaBoost	6000	300	0.9	No

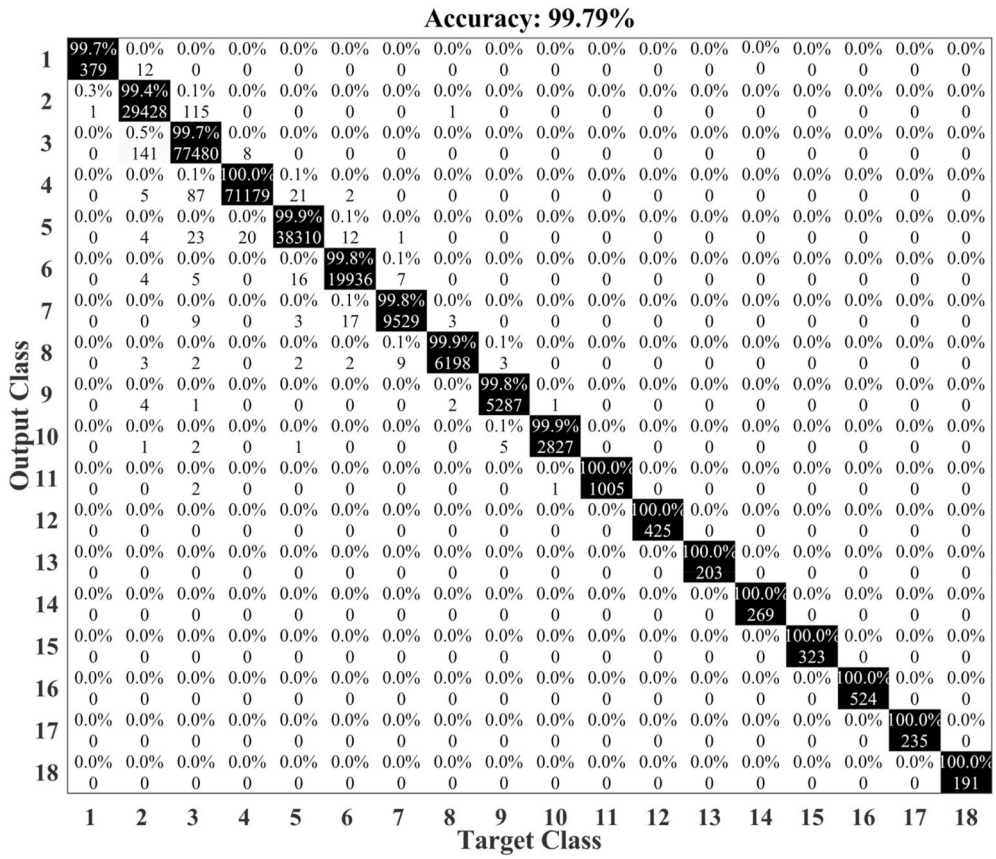

Fig. 11. Confusion matrix for the case of flat classification on the training dataset.

of the hierarchy. Fig. 7 demonstrates that at the second level of the hierarchy, class overlapping is also minimized. The above observations suggest that balancing the hierarchically decomposed dataset via SMOTE of the minority classes or cropping the majority class would be more appropriate through the framework of hierarchical decomposition.

The authors are aware of the limited literature on hierarchical, machine-learning based classification methods in the wave prediction. Still, many studies in different application domains suggest that hierarchical methods can have better classification performance compared to flat classification techniques (Epshtein and Uliman, 2005; Freeman et al., 2011; Hao et al., 2007; Huang et al., 2013). For a comprehensive review on the combination of different hierarchical classifier in different application domains, the reader is referred to the study of Silla and Freitas (2011).

2.3. Hierarchical decomposition of significant wave height

Unlike other areas of machine learning classification where intuitive hierarchical decomposition of the classes can be performed e.g., the classification of animal species into genus, family, order, class etc. The case of significant wave height classification is different in the sense that no prior knowledge of the hierarchical structure exists. As a result, while it is possible to divide the succeeding levels in the wave height hierarchy in arbitrary labeled classes based on a general understanding of the different wave height ranges, for example, low, medium, large, extreme etc., the authors believe that a structured approach based on the clustering of data in the feature space can help in identifying the natural occurring significant wave height hierarchy.

In this study, this natural decomposition of the significant wave height is performed systematically through an unsupervised clustering technique, namely; the weighted pair group method with averaging (WPGMA) (Sokal and Michener, 1958). This simple agglomerative bottom-up hierarchical clustering method allows the clustering of features based on their similarity. In this regard, the center of mass of each of the 28 features described in section 2 is identified and fed to the WPGMA. The distance (linkage) between each adjacent feature is calculated through the norm metric and clustered accordingly. Fig. 8 summarizes the results of the WPGMA clustering method, exhibiting the naturally occurring clustering of significant wave height classes.

With reference to Fig. 8 following a top-bottom decomposition the following binary clusters are identified, namely: Binary cluster 1 distinguishing between wave ranges [0, 2.4) and [2.4, 3.6], Binary cluster 2 distinguishing between wave ranges [0, 0.2) and [0.2, 2.4) and Binary cluster 3 distinguishing between wave ranges [0.2, 0.8) and [0.8, 2.4). Subsequently, three multiclass clusters are identified, namely: Multiclass cluster 1 distinguishing between wave ranges [0.2, 0.4), [0.4, 0.6) and [0.6, 0.8), Multiclass cluster 2 distinguishing between wave ranges [0.8, 1.0), [1.0, 1.2) …, [2.2, 2.4) e.g. class IDs (6–12), and finally Multiclass cluster 3 distinguishing between wave ranges [2.4, 2.6), [2.6, 2.8) …,[3.4, 3.6] e.g. class IDs (13–18). The reader is referred back to Fig. 5 for a schematic representation of this particular hierarchical decomposition.

Accordingly, hierarchical classification based on the presented decomposition entails the training of six independent classifiers of which three are binary classifiers and the remaining three multiclass classifiers. While it is possible to further decompose the problem in more binary local classifiers, it can be understood at lower levels of the hierarchy, feature dissimilarity decreases making the distinction of such clusters difficult. Seemingly the investigation of the performance of hierarchical classification at varying levels of decomposition is an interesting aspect to be considered. Still, such an investigation deviates from the initial hypothesis and scope of the present study in examining the performance of hierarchical over flat classification methods.

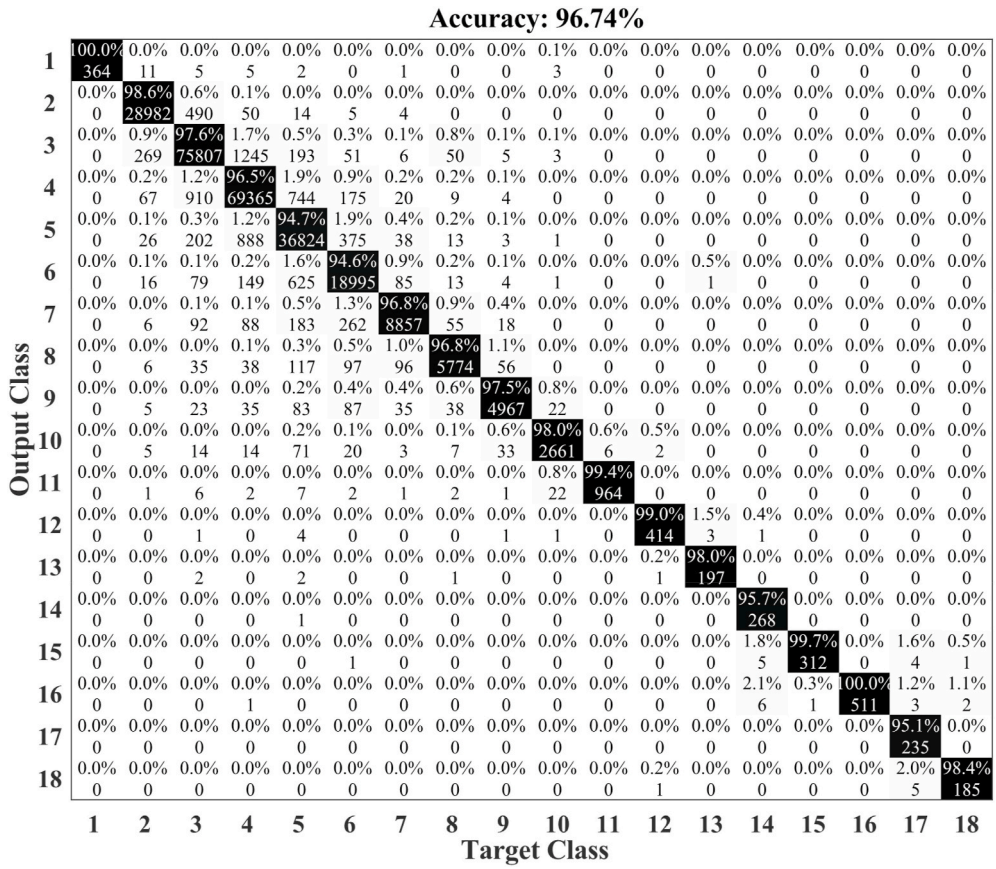

Fig. 12. Confusion matrix for the case of hierarchical classification on the training dataset.

2.4. Training of flat and hierarchical classifiers

In accordance to good machine learning practice, the dataset was divided in two groups for the respective training and testing of the developed classifiers on independent datasets. Accordingly, the first set consisting of 85% of wind, wave and structural data was used for training the classifiers. The remaining 15% of the dataset was retained as a testing dataset for providing an unbiased evaluation of the final flat and hierarchical classifiers. Fig. 9, demonstrates the distribution of samples across the training and testing datasets exhibiting the proportional distribution of instances per class in both datasets.

During training, a 5-fold cross-validation procedure was adopted. Accordingly, at each complete training cycle, a total of $K-1$ e.g. 4 different training attempts were made. At each training attempt, every classifier was trained on the $K-1$ fold data while the first fold was treated as the validation set. In doing so, each sample in the dataset was given the opportunity to be used in the holdout set 1 time while also used to train the model $K-1$ times.

For the training of the three binary classifiers identified from the natural hierarchical decomposition of significant wave heights described in the preceding section, SMOTE was performed on the training dataset to increase the number of samples in the minority class. Similarly, under-sampling of the majority class was performed on the training dataset by removing random samples from the majority class. Both the oversampling ratio in SMOTE and the under-sampling ratio were selected through Global Search (Zsolt et al., 2007) gradient based nonlinear optimization with multiple start points so that a global maximum is achieved. The objective function in the optimization routine was described by the macro average Fscore achieved by a degenerate (e.g. non-optimized) K-nearest neighbor classifier on the testing dataset attributed to each binary cluster. An example of the optimization results obtained for binary cluster 3 (e.g. [0.2, 0.8) and [0.8, 2.4)) is presented in Fig. 10. With reference to Fig. 10, it is observed that maximum classification performance is achieved when the minority class is oversampled by ~17% of synthetic data, while at the same time the majority class is under-sampled by ~5%. Similar procedure has been adopted for the remaining two binary clusters. The optimized oversampling and under-sampling ratios are summarized in Table 2.

A step-by-step optimization procedure was adopted for determining the hyperparameters for both the binary and multiclass classifiers in accordance to Demetriou et al. (2021). Post-optimization, feature curtailment was performed by identifying the least relevant features through the Gini Impurity index. For brevity, the reader is referred to Demetriou et al. (2021) for a comprehensive description of the optimization and feature curtailment procedure adopted herein. The results of the optimization are presented in Table 2 below. It is noteworthy that under-sampling and SMOTE for the case of multiclass classifiers was not performed owing to significant data overlap as described in the previous sections. Nevertheless, it was deemed appropriate to focus optimization efforts on the binary classifiers as slight deviations in accuracy and corresponding classification errors will propagate down the hierarchy.

Table 3 summarizes the time required to complete the training cycles of both flat and hierarchical classifiers. The size of data used in the training of each classifier after SMOTE and feature curtailment is also exhibited on Table 3. It is worth noting that despite the small increase in data size between the flat classifier (e.g. 65 MB) and the hierarchical binary classifier utilizing the largest amount of data in training (e.g. 54 MB), a significant increase in training times for the case of flat classification is observed. This increase is attributed primarily to the number of learners required in the training of the flat classifier as identified from the optimization procedure. Aggregating the total training times of flat and hierarchical classifiers, expending a total of 2138.9s and 1305.9s respectively, a reduction of 38.95% in training times is observed in favor of the latter.

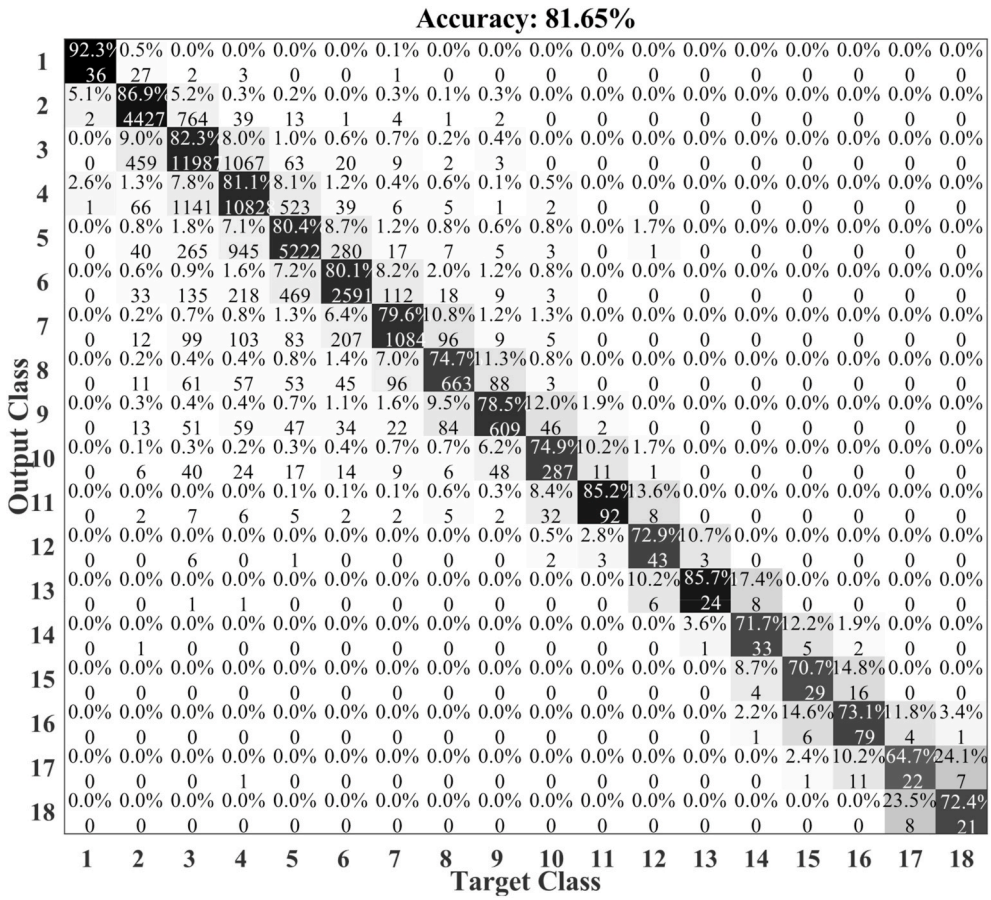

Fig. 13. Confusion matrix for the case of flat classification on the testing dataset.

3. Performance comparison of flat vs hierarchical classifier

Proving the initial hypothesis and demonstrating that hierarchical decomposition can aid the prediction of significant wave heights necessitates the thorough comparison of flat over hierarchical classifiers. In this regard, a superior methodology should yield a model that ultimately is capable of distinguishing between 18 classes of significant wave heights (e.g. 0–3.6 m in 0.2 m bins) at improved classification accuracy. To examine this hypothesis, a series of performance metrics that take into account bias in the case of imbalance datasets are established based on the following four classifier-response categories:

- **True Positive (TP):** The model correctly predicts the positive class
- **True Negative (TN):** The model correctly predicts the negative class
- **False Positive (FP):** The model incorrectly predicts the positive class
- **False Negative (FN):** The model incorrectly predicts the negative class

In the above, the definition positive class is attributed to the particular class of interest e.g. if the objective is to identify class 1, this class attains the role of the positive class, while the remaining classes (e.g. class 2–18) are considered to be negative. Similarly, when class 2 is to be detected, class 2 becomes the positive class while the remaining classes e.g. 1 and 3–18 are considered the negative classes.

From these, common multi-class evaluation metrics for machine learning classification models are established and calculated according to the following metrics:

Accuracy: Number of observations identified to be either truly positive or truly negative out of the total number of observations.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

It is worth highlighting that since the classes are not equally represented, the accuracy metric might not be representative of the actual classifier's performance. This can be exemplified by assuming a degenerate machine learning classifier that assigns class 2 to all the observations. Since class 2 represents approximately 25% of the observations, the classifier attains an accuracy of 25% with no classification effort. For this reason, while the accuracy metric is retained, other objective functions, namely, Fscore, Informedness and Markedness will be later defined, as they are considered more appropriate measures of classification accuracy, particularly for the cases of multi-class classification problems with imbalanced datasets. To calculate these objective functions, the following expressions are defined:

True positive rate - TPR (sensitivity/recall): Number of observations correctly identified as positive out of the total actual positives

$$TPR = \frac{TP}{(TP + FN)} \tag{2}$$

Positive predictive values (PPV) (precision): Number of observations correctly identified as positive out of the total items identified as positive

$$PPV = \frac{TP}{(TP + FP)} \tag{3}$$

True negative rate (TNR) (specificity): Number of observations correctly identified as negative out of the total actual negatives

$$TNR = \frac{TN}{(TN + FP)} \tag{4}$$

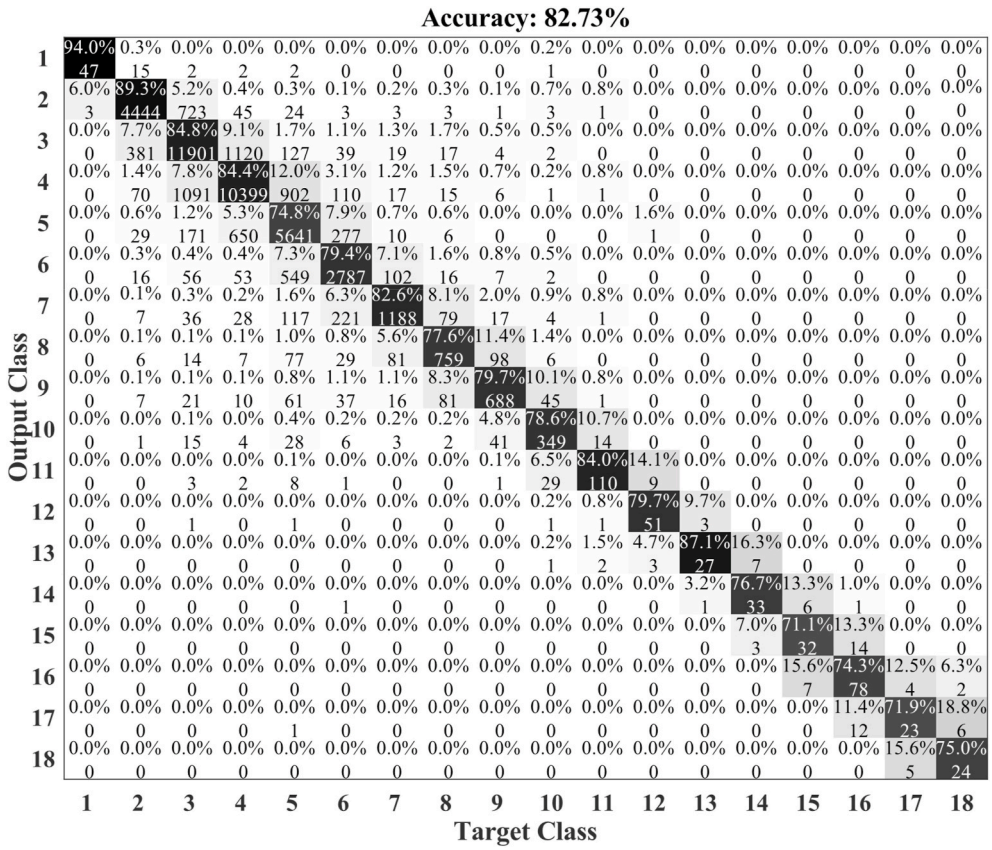

Fig. 14. Confusion matrix for the case of hierarchical classification on the testing dataset.

Fscore: The harmonic average of the precision and recall, it measures the effectiveness of identification when just as much importance is given to sensitivity as to precision

$$Fscore = \frac{2TPR \times PPV}{(TPR + PPV)} \tag{5}$$

To examine the performance of the developed flat and hierarchical classifiers, each type of classifier was exposed to the training and testing dataset and each of the responses (observations) was recorded on a confusion matrix. In this regard, Fig. 11 Fig. 12, Fig. 13 and Fig. 14 summarise the response of the developed flat and hierarchical classifiers respectively in the form of a confusion matrix, along with a summary of the macro accuracy calculated with Equation (1).

With reference to Figs. 11, Fig. 12, Figs. 13 and 14, the number of observations along the diagonal correspond to the true positive (TP) responses for each of the 18 classes. The row-wise summation of the observations corresponds to the false positive (FP) responses attributed to the class in consideration. Similarly, the column-wise summation of the observations corresponds to the false negative (FN) responses attributed to the particular class. Finally, true negative (TN) responses of a particular class are all the observed instances excluding the TP, FP and FN values for the class in consideration. From the above recorded classifier responses and with reference to Figs. 11 and 12, the metric of Accuracy established in Equation (1) is calculated as 99.79% and 96.74% for the case of flat and hierarchical classification on the training dataset. The higher Accuracy attributed to flat classification suggests that: 1) the flat classification model is overfitted on the training data and 2) the Accuracy metric does a poor job in representing the actual performance of classifiers on cases with imbalanced data. Accordingly, the unbiased investigation of the performance of both the flat and hierarchical classifiers necessitates the thorough investigation of their responses on the testing dataset which has remained unseen by both models during training, as well as the establishment of further performance metrics complementary to Accuracy. With reference to Figs. 13 and 14, the Accuracy on the testing dataset is calculated as 81.65% and 82.73% for the case of flat and hierarchical classification respectively. Upon initial observation, the single percentage (~1%) performance improvement validates the initial assumptions that 1) the flat classification models was overfitted on the training dataset and 2) the Accuracy metric does a poor job in representing the actual classification performance for imbalanced dataset cases.

To address the issue of the Accuracy metric not being representative for the case of class imbalance, the metrics of TPR, PPV, and their derivative Fscore were calculated in accordance to Equation (2), Equation (3) and Equation (5). Fig. 15 summarizes the achieved TPR, PPV, and Fscore for the flat and hierarchical classifiers on the testing dataset, demonstrating the superiority of the hierarchical over the flat classification approach across the majority of the classes.

Similar procedure is adopted for establishing metrics that account for chance or base case level the statistics in the dataset. Accordingly, the metrics of Informedness, and Markedness (Powers, 2007) are established through the following expressions.

Informedness: Measures how informed is the network about both positives and negatives according to the following expression:

$$Informedness = \frac{TP}{RP} - \frac{FP}{RN} \tag{6}$$

In which $RP = TP + FN$ is the total actual positives and $RN = TN + FP$ represents the total actual negatives.

Markedness: Measures the trustworthiness of positive and negative predictions by the network and is expressed by the following.

$$Markedness = \frac{TP}{PP} - \frac{FN}{PN} \tag{7}$$

In which $PP = TP + FP$ is the total number of observations identified

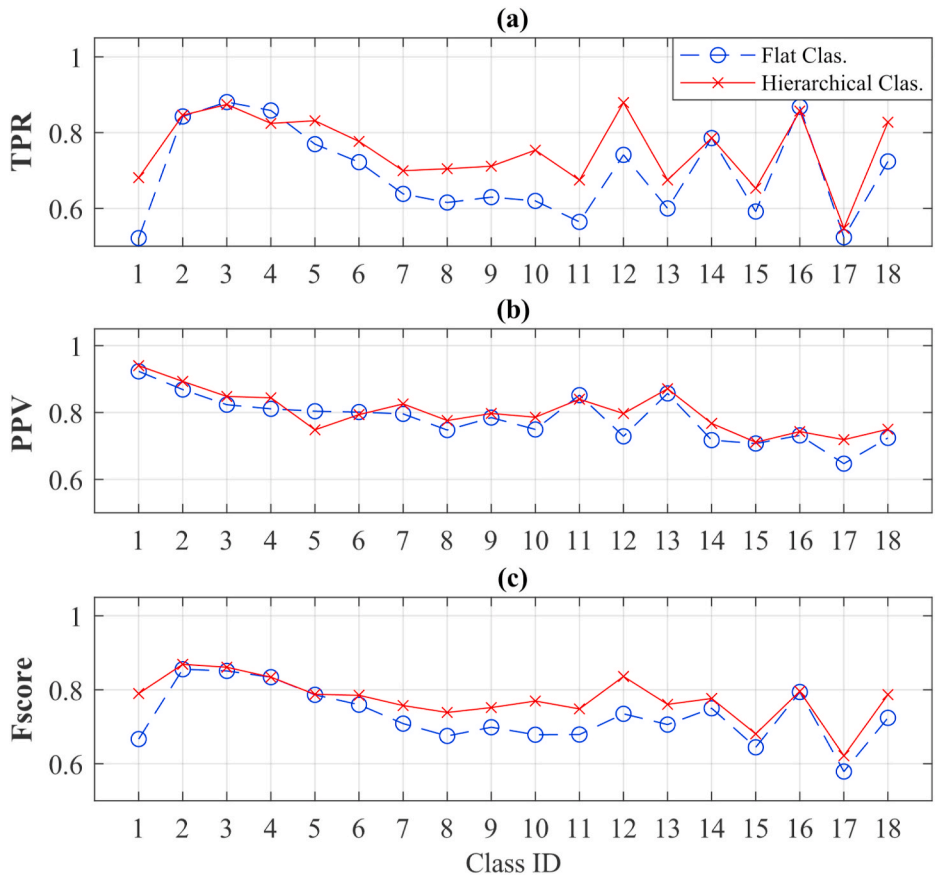

Fig. 15. Performance comparison of flat vs hierarchical classifiers based on the metrics of (a) TPR, (b) PPV, and (c) Fscore.

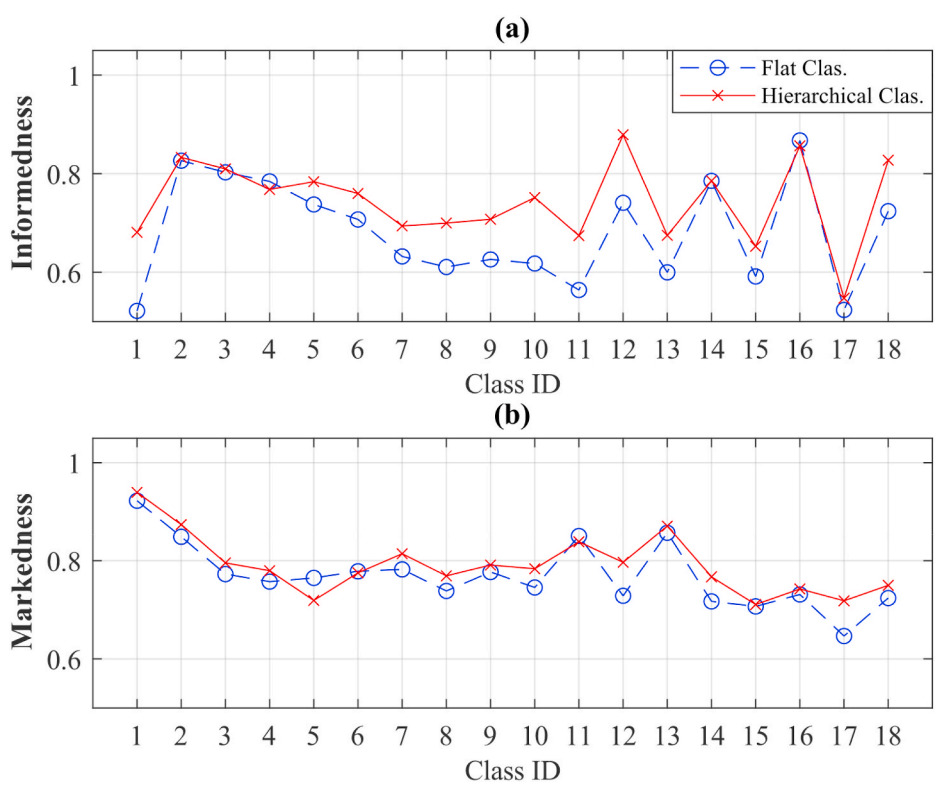

Fig. 16. Performance comparison of flat vs hierarchical classifiers based on the metrics of (a) Informedness, and (b) Markedness.

Table 4
Macro-Average Performance Metrics and total computational time.

Class ID	Performance Metric						Total Computational Time (s)
	Accuracy	TPR	PPV	Fscore	Informedness	Markedness	
Flat Classifier	0.8165	0.6944	0.7819	0.7292	0.6813	0.7694	17.17
Hierarchical Classifier	0.8273	0.7558	0.8028	0.7751	0.7437	0.7909	11.90

as positives and $PN = FN + TN$ represent the total number of observations identified as negatives.

From the above expressions, it is evident that while TPR, PPV and Fscore are only dependent on the TP, FP, FN, RP, PP remain unaffected by the changes to the values of TN, RN and PN. On the contrary, Informedness and Markedness are a function of all the above, suggesting that these metrics being more sensitive to changes in the classifier's performance. Nevertheless, each metric serves different purposes. Still, consideration of a collection of metrics allows a more comprehensive study of classification performance. Fig. 16 summarizes the achieved Informedness and Markedness scores for the flat and hierarchical classifiers.

For completeness, the macro-average of the metrics of Accuracy, TPR, PPV, Fscore, Informedness, and Markedness (e.g. the mean scores of all classes) is calculated and the results are exhibited in Table 4 below. The total time required by the flat and hierarchical classifiers to perform significant wave height predictions on the testing dataset is also presented in Table 4.

4. Conclusions

This study proposes the hierarchical decomposition of significant wave heights for improving the prediction accuracy of multiclass classification models. The study is founded on the idea that through the framework of hierarchical decomposition, classification of imbalanced datasets can be improved owing to the reduced class imbalance ratio between the minority and majority classes at each level of the hierarchy. At the same time, it was hypothesized that reducing the classification problem to a collection of binary and multiclass classifiers, flexibility is achieved in optimizing each of the local classifiers with respect to the sampling ratios in the SMOTE and under-sampling of data pre-training, selecting appropriate algorithm and optimizing its hyperparameters during training, and finally performing feature curtailment post-training. In examining this hypothesis, data extracted from the structural and meteorological monitoring station installed on the coastal jetty located in the southern coasts of Cyprus were used in the development of flat and hierarchical classifiers. In this regard, the single flat classifier was developed based on the procedure described extensively by the authors in Demetriou et al., (2021), while the hierarchical development approach consisted of dividing the feature space in clusters via the WPGMA method for hierarchical clustering. Accordingly, three binary and three multiclass classifiers were developed based on the identified clusters. Each of the three binary classifiers was trained on a SMOTE adjusted/under-sampled training dataset for improving its classification performance on the testing dataset. Subsequently, each of the binary and multiclass classifier's algorithm, hyperparameters and number of features was optimized based on the procedure followed in the case of flat classification. It is worth highlighting that for both the training of flat and hierarchical classifiers, measures were taken for the development of unbiased models using procedures such as train-test splits and K-fold validation. Post-training, a variety of performance metrics were established for reflecting the probability that the model's prediction is informed rather than being based on chance. It has been demonstrated that despite the complexity arising from the performance of various operations such as WPGMA, and SMOTE as well as the apparent complexity arising from the development of various classifiers contrary to a single flat one, the hierarchical approach was shown to outperform the single classification approach with respect to the established metrics as well as the times required in training the models and the time expensed for performing significant wave height predictions.

The findings of this study highlight the great promise of hierarchical classification of significant wave heights as well as the potential for future work to be undertaken in this direction. Finally, it is important to highlight that the apparent complexity associated with developing hierarchical classification models necessitates the development of a rigid framework for the development of such models.

Funding

This work was supported by the Ministry of Energy Commerce and Industry of Cyprus; the Cyprus University of Technology; and VTT Vassiliko Ltd.

CRediT authorship contribution statement

Demetris Demetriou: Conceptualization, Methodology, Investigation, Software, Formal analysis, Data curation, Writing – original draft. **Constantine Michailides:** Conceptualization, Methodology, Funding acquisition, Investigation, Writing – review & editing, Project administration. **George Papanastasiou:** Resources. **Toula Onoufriou:** Conceptualization, Methodology, Funding acquisition, Investigation, Supervision, Validation, Writing – review & editing, Project administration.

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.

Acknowledgments

The authors gratefully acknowledge the Ministry of Energy, Commerce, and Industry of Cyprus and the Cyprus University of Technology for their financial support to this study. Additionally, the authors gratefully acknowledge the support of VTT Vassiliko Ltd. who provided access to their facility and data.

References

Bergin, W., Spearman, E., 2006. Vasilikos Energy Centre Basis of Design Environmental Assessment. Parsons Brinckerhoff Limited.
Beyan, C., Fisher, R., 2015. Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recogn. 48 (5), 1653–1672.
Bolton, S., 2014. In: Chapter Thirteen - Operation and Maintenance, Offshore Wind, second ed. Academic Press, pp. 243–283.
Browne, M., Castelle, B., Strauss, D., Tomlinson, R., Blumenstein, M., 2007. Near-shore swell estimation from a global wind–wave model: spectral process, linear, and artificial neural network models. Coast. Eng. 54 (5), 445–460.
Buda, M., Maki, A., Mazurowski, M., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Network. 106, 249–259.
Ceci, M., Malerba, D., 2007. Classifying Web documents in a hierarchy of categories: a comprehensive study. J. Intell. Inf. Syst. 28 (1), 37–78.
Chawla, N., Lazarevic, A., Hall, L.O., Bowyer, K., 2003. SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (Eds.), Knowledge Discovery in Databases: PKDD 2003. PKDD 2003. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Das, B., Krishnan, N., Cook, D., 2014. Handling imbalanced and overlapping classes in smart environments prompting dataset. In: Yada, K. (Ed.), Data Mining for Service. Studies in Big Data. Springer, Berlin, Heidelberg.
Demetriou, D., Michailides, C., Papanastasiou, G., Onoufriou, T., 2021. Coastal zone significant wave height prediction by supervised machine learning classification algorithms. Ocean Eng. 221.
Deo, M.C., Jha, A., Chaphekar, A.S., Ravikant, K., 2001. Neural networks for wave forecasting. Ocean Eng. 28, 889–898.
Egozcue, J.J., Pawlowsky-GlahnVera, V., Ortego, M.I., 2005. Wave-height hazard analysis in Eastern Coast of Spain–Bayesian approach using generalized Pareto distribution. Adv. Geosci. 2, 25–30.
Epshtein, B., Uliman, S., 2005. Feature hierarchies for object classification. In: Tenth IEEE International Conference on Computer Vision (ICCV'05), pp. 220–227.
Freeman, C., Kulić, D., Basir, O., 2011. Joint feature selection and hierarchical classifier design. IEEE Int. Conf. Syst. Man Cybern. 1728–1734.
García, S., Herrera, F., 2009. Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy, Evolutionary Computation. MIT Press, pp. 275–306.
Group, T.W., 1988. The WAM model—a third generation ocean wave prediction model. J. Phys. Oceanogr. 18 (12), 1775–1810.
Gunaydın, K., 2008. The estimation of monthly mean significant wave heights by using artificial neural network and regression methods. Ocean Eng. 35, 1406–1415.
Han, H., Wang, W., Mao, B., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D., Zhang, X., Huang, G. (Eds.), Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.
Hao, P., Chiang, J., Tu, Y., 2007. Hierarchically SVM classification based on support vector clustering method and its application to document categorization. Expert Syst. Appl. 33 (3), 627–635.
Hasselmann, K., Barnett, T.P., Bouws, E., Carlson, H., Cartwright, D.E., Enke, K., Ewing, J.A., Gienapp, H., Hasselmann, D.E., Kruseman, P., Meerburg, A., Mller, P., Olbers, D.J., Ritcher, K., Sell, W., Walden, H., 1973. Measurements of Wind–Wave Growth and Swell Decay during the Joint North Sea Wave Project (JONSWAP), Ergänzungsheft 8-12. Deutches Hydrographisches Institut.
Hersbach, H., Janssen, P., 1999. Improvement of the short-fetch behavior in the wave ocean model (WAM). J. Atmos. Ocean. Technol. 16, 884–892.
Huang, P., Boom, B., Fisher, R., 2013. Underwater live fish recognition using a balance-guaranteed optimized tree. In: Lee, K., Matsushita, Y., Rehg, J., Hu, Z. (Eds.), Computer Vision – ACCV 2012. Springer, Berlin, Heidelberg.
Japkowicz, N., Stephen, S.p., 2002. The Class Imbalance Problem: A Systematic Study Intelligent Data Analysis, pp. 429–449.
Koller, D., Sahami, M., 1997. Hierarchically classifying documents using very few words. In: Fourteenth Internation Conference on Machine Learning (ML-97), pp. 170–178. Nashville, Tenessee.
Kumar, N.K., Savitha, R., Mamun, A., 2018. Ocean wave height prediction using ensemble of Extreme Learning Machine. Neurocomputing 277, 12–20.
Li, M., Liu, K., 2020. Probabilistic prediction of significant wave height using dynamic bayesian network and information flow. Water 12 (9).
Mahjoobi, J., Mosabbeb, E.A., 2009. Prediction of significant wave height using regressive support vector machines. Ocean Eng. 339–347.
Mahjoobi, J., Shahidi, A.I., Kazeminezhad, M.H., 2008. Hindcasting of wave parameters using different soft computing methods. Appl. Ocean Res. 30 (1), 28–36.
Malakar, S., Ghosh, M., Bhowmik, S., Sarkar, R., Nasipuri, M., 2020. A GA based hierarchical feature selection approach for handwritten word recognition. Neural Comput. Appl. 32, 2533–2552.
Mentaschi, L., Besio, G., Cassola, F., Mazzino, A., 2015. Performance evaluation of wavewatch III in the Mediterranean sea. Ocean Model. 90, 82–94.
Monbaliu, J., Padilla-Hernández, R., Hargreaves, J.C., Albiach, J., Luo, W., Sclavo, M., Günther, H., 2000. The spectral wave model, WAM, adapted for applications with high spatial resolution. Coast. Eng. 41, 41–62.
Ou, S.H., Liau, J.M., Hsu, T.W., Tzang, S.Y., 2002. Simulating typhoon waves by SWAN wave Model in Coastal Waters of Taiwan. Ocean Eng. 29, 947–971.
Powers, D., 2007. Evaluation: from Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001. School of Informatics and Engineering Flinders University of South Australia.
Rao, S., Mandal, S., 2005. Hindcasting of storm waves using neural networks. Ocean Eng. 32 (5–6), 667–684.
Rogers, W.E., Kaihatu, J.M., Hsu, L., Jensen, R.E., Dykes, J.D., Holland, K.T., 2007. Forecasting and hindcasting waves with the SWAN model in the southern California. Bight Coast. Eng. 54 (1), 1–15.
Scotto, M.G., Soares, C.G., 2007. Bayesian inference for long-term prediction of significant wave height. Coast. Eng. 54, 393–400.
Secker, A., Davies, M., Freitas, A., Clark, E., Timmis, J., Flower, D., 2010. Hierarchical classification of G-Protein-Coupled Receptors with data-driven selection of attributes and classifiers. Int. J. Data Min. Bioinf. 4 (2).
Seiffert, C., Khoshgoftaar, T., Hulse, J., Napolitano, A., 2008. RUSBoost: improving classification performance when training data is skewed. In: 19th International Conference on Pattern Recognition, pp. 1–4.
Shamshirband, S., Mosavi, A., Rabczuk, T., Nabipour, N., 2020. Prediction of significant wave height; comparison between nested grid numerical model, and machine learning models of artificial neural networks, extreme learning and support vector machines. Eng. Appl. Comput. Fluid Mech. 14 (1), 805–817.
Sharma, N., Jain, P., 2018. A particle swarm optimization based machine learning approach for wave height prediction. Int. J. Innovat. Res. Comput. Commun. Eng. 6 (7).
Silla, C., Freitas, A., 2011. A survey of hierarchical classification across different application domains. Data Min. Knowl. Discov. 22 (1), 31–72.
Sokal, R., Michener, C., 1958. A Statistical Method for Evaluating Systematic Relationships. University of Kansas Science Bulletin, pp. 1409–1438.
Takerkhani, A., Cosma, G., McGinnity, T., 2020. AdaBoost-CNN: an adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. Neurocomputing 404, 351–366.
Tang, Y., Zhang, Y., Chawla, N., Krasser, S., 2009. SVMs modeling for highly imbalanced classification. IEEE Trans. Syst. Man Cybern. B (Cybernetics) 281–288.
Tolman, H.L., Krasnopolsky, V., Chalikov, D., 2005. Neural network approximations for nonlinear interactions in wind wave spectra: direct mapping for wind seas in deep water. Ocean Model. 8 (3), 253–278.
Wang, J., Wang, Y., Yang, J., 2021. Forecasting of significant wave height based on gated recurrent unit network in the Taiwan strait and its adjacent waters. Water 13 (1).
Wei, C.-C., 2018. Nearshore wave predictions using data MiningTechniques during typhoons: a case study near Taiwan's northeastern coast. Energies 11 (11).
Wu, M., Stefanakos, C., Gao, Z., Haver, S., 2019. Prediction of short-term wind and wave conditions for marine operations using a multi-step-ahead decomposition-ANFIS model and quantification of its uncertainty. Ocean Eng. 188.
Zheng, W., Zhao, H., 2020. Cost-sensitive hierarchical classification for imbalance classes. Appl. Intell. 50, 2328–2338.
Zijlema, M., 2010. Computation of wind–wave spectra in coastal waters with swan on unstructured grids. Coast. Eng. 57 (3), 267–277.
Zsolt, U., Lasdon, L., Plummer, J., Glover, F., Kelly, J., Martí, R., 2007. Scatter Search and local NLP solvers: a Multistart framework for global optimization. Inf. J. Comput. 19 (3), 328–340.