Wave Characteristics Prediction Using Raw Data and Multi-Output Machine Learning Algorithms: Towards a Data-Driven Wave Energy System Development

Masoud Masoumi*

Department of Mechanical Engineering, Manhattan College, Riverdale, New York, 10471, USA

Abstract

This work introduces the use of multi-output regression algorithm for wave height and wave period prediction in the United States waters using the recorded data from 104 stations, from 2010 to 2019. The models use raw data for all the stations monitored by National Oceanic and Atmospheric Administration's National Data Buoy Center. Five models are developed using four machine learning algorithms of K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Regression (SVR), and Neural Network (NN). These models take a latitude, a longitude, and a month as inputs and predict three features, which are monthly maximum, monthly average and monthly minimum values for wave height and wave period at the given location in a given month. Results showed that the models developed based on DT, KNN, and NN algorithms have good performances, especially in terms of the monthly minimum and monthly average value prediction for both wave height and wave period values.

Keywords: Wave Prediction, Ocean Wave Characteristics, Multi-Output Regression, Data-Driven Modeling, Machine Learning

1. Introduction

Looking at the vast and untapped energy of the ocean waves, researchers and engineers have been developing various machines and technologies to harness this energy and convert it into electricity. There are multiple challenges involved in designing and installing a wave energy converter - salinity of the water in the ocean, survivability of the device under extreme weather conditions, power transmission to the land, power take-off mechanisms that should have high efficiency under various wave characteristics to name a few. The work in this study is concerned with the last challenge, which is the development of an effective power take-off mechanism capable of efficiently converting wave energy into electricity for waves with various characteristics. To that end, one important step is to develop an easy-to-implement wave characteristics prediction model that can be integrated into the wave energy converter design procedure.

Numerous models have been developed and used for wave height and wave period prediction at a given location based on the historical data in that location or its nearby sites. These models can generally be classified into four categories, including first, second, third, and improved third generation wave models (Mandal & Prabaharan, 2010). All these modeling approaches are based on the spectral energy-balance equation (Kumar, Savitha, & Al Mamun, 2018).

Aside from the models developed based on the energy-balanced equation, there are models constructed using machine learning techniques, which are not limited by the simplifications involved in the four categories of models mentioned above. A prominent machine learning technique used for wave forecasting is Neural Network. An overview of the ocean wave characteristics forecasting using Neural Network (NN) can be found in Mandal and Prabaharan's work (Mandal & Prabaharan, 2010).

While different types of NNs have been used for wave characteristics prediction and forecasting, there are also works

* Received Date: 2021.11.30
E-mail of corresponding author: mmasoumi01@manhattan.edu

focused on using other machine learning techniques such as Support Vector Machine (Berbić, Ocvirk, Carević, & Lončar, 2017), K-Nearest Neighbors (Nikoo, Kerachian, & Alizadeh, 2018), Bayesian Networks and Adaptive Neuro-Fuzzy Inference System (Malekmohamadi, Bazargan-Lari, Kerachian, Nikoo, & Fallahnia, 2011), Sequential Learning Neural Networks (Savitha, Al Mamun, & others, 2017), Extreme Learning Machines (Ali & Prasad, 2019), Support Vector Regression, Deep Belief Networks (Kumar, Savitha, & Al Mamun, 2018), and Recurrent Neural Network (Pirhooshyaran & Snyder, 2020).

Berbic et al. used Support Vector Machine (SVM) to predict significant wave height based on the data from two waverider stations, recorded from November 2007 to November 2008 (Berbić, Ocvirk, Carević, & Lončar, 2017). Further, K-Nearest Neighbors (KNN) was implemented to develop a model to forecast significant wave height based on the wave and wind time series data, recorded at station 45006 (located at Ironwood, MI) from March 2005 to December 2006 (Nikoo, Kerachian, & Alizadeh, 2018). Moreover, the data recorded from the Caspian Sea from 2006 was utilized to develop a forecasting model using KNN algorithm (Zamani, Solomatine, Azimian, & Heemink, 2008). Implementing Bayesian Networks and Adaptive Neuro-Fuzzy Inference, Malekmohamadi et al. used the wave height data from western part of the Lake Superior during 2006 to develop a wave forecasting model (Malekmohamadi, Bazargan-Lari, Kerachian, Nikoo, & Fallahnia, 2011). Focusing on the UK, Korean region, and the Gulf of Mexico, Savitha et al. implemented sequential learning neural networks (Savitha, Al Mamun, & others, 2017) as well as an ensemble of extreme learning machines (N. Krishna Kumar, 2018) to make daily prediction on wave height. The data were collected by four buoys in the Guld of Mexico, four buoys in Korean region, and five buoys in the UK for the first work. For the second work, the data were collected by 10 stations in the Gulf of Mexico, Brazil, and Korean region.

Using the data from 2011 to 2014, Kumar et al. (Kumar, Savitha, & Al Mamun, 2018) developed three single output Deep Belief Networks, each predicting one wave characteristic, including significant wave height, peak wave period, and zero crossing period. They used the data from the UK (four stations), Korean region (four stations), the Gulf of Mexico (four stations), and Irish region (five stations). Root mean square error and correlation coefficient were implemented as evaluation metrics. The model was used to predict sea state characteristics and then the results were implemented to calculate the wave-drift force and moment exerted on marine structures. They also compared the outcomes with the outcomes of an extreme machine learning, a support vector regression, and an online sequential extreme learning. A time series of the data for 2015 from Italy along with the data from 2012 Global Energy Forecasting Competition were used to develop an ensemble of wavenet learners by Ribeiro et al. (Ribeiro, Mariani, & dos Santos Coelho, 2019). They implemented a Box-Cox transformation to normalize the data and remove the trends in the dataset. Then, a set of selected features were used to train the model.

Ali and Prasad used the historical time data of the Gold Coast and Mooloolaba in Australia from 2000 to 2018 to propose a machine learning model built using a combination of extreme learning machines and empirical mode decomposition. They trained the model on 70% of the data and test its performance, in terms of predicting the significant wave height, using 30% of the data (Ali & Prasad, 2019). Both recurrent and sequence-to-sequence networks were also used to predict wave characteristics (Pirhooshyaran & Snyder, 2020). They used the data collected from a buoy monitored by the National Oceanic and Atmospheric Administration's National Data Buoy Center to predict the significant wave height. Karabulut and Ozmen Koca proposed a model that took the month, day, and flow velocity at different water depth and could make predictions on significant wave height values (Karabulut & Ozmen Koca, 2020). Their model was train and tested on the data from two different buoys located in the south coast of Turkey.

The machine learning algorithms used to make predictions on and to forecast both the wave height and wave period have been mainly focused on achieving high accuracy using the historical data from one or a few buoy stations. These models are typically using time series data and are trained to predict significant wave height or wave period for the location, where the data was collected. Granted these models perform well and provide good accuracy specially for short term forecasting, the goal in this work is to develop simple and easy-to-implement machine learning models to estimate the wave characteristics for a given location based on the data from a large network of buoys. The main difference between this work and previously published works is that this work is focused on wave height and wave period prediction for any location in the US coastal region for a given month. The models developed in this work can estimate monthly maximum, monthly minimum, and monthly average wave height and wave period based on ten years of data collected by 104 stations monitored by National Data Buoy Center. This modeling approach is adopted to provide a platform that can specially be used by wave energy converter developers who need to have some initial estimates of wave height and wave period ranges for a possible installation site for their wave energy system. This can help them design their device for maximum efficiency for a given location. It also provides them with a tool to explore optimal control systems, which can immensely help maximize power generation during each month of the year (Drew, Plummer, & Sahinkaya, 2009; Li, Weiss, Mueller, Townley, & Belmont, 2012).

The rest of this report is organized as follows. Section 2 focuses on the machine learning algorithms that are used to develop the models along with two evaluation metrics. Section 3 provides the results for the model development and the performance of each model in terms of each target variable i.e., monthly minimum, monthly maximum, and monthly average wave height and wave period. The performances of the models are compared, and the results are discussed. Section 4 provides the final comments and conclusions.

2. Methods

In this section, four machine learning techniques, used to develop multi-output regression models, are briefly reviewed. Further, the evaluation metrics that are used to assess the performance of these models are also discussed. References are provided for further information regarding each topic. These algorithms include K-Nearest Neighbors, Decision Tree, Support Vector Regression, and Neural Network. In all these models, the inputs are the latitude and longitude of the location as well as the month of the year. The outputs are maximum, minimum, and average values for either wave height or wave period. All models in this work are trained on 80% and tested on 20% of the data unless otherwise stated.

2.1 K-Nearest Neighbors (KNN)

When using KNN technique for regression, the goal is to learn a regression function that can predict the outcomes for new data points based on a set of K nearest observations to that data point. Given an input point $x^{(p)}$ and a dataset with m points, KNN algorithm first finds the distances between the given point and all the points in the dataset. Afterwards, it predicts the target variable for the given point based on the values of target variables of the nearby points (Cover, 1968; Song, Liang, Lu, & Zhao, 2017). The tunning parameter for KNN technique is the K value, number of nearby points that are used to make prediction for the target variable. This parameter can be found either through running the model using different values of K and evaluating its performance or through a cross-validation approach (Kramer, 2013).

2.2 Decision Tree (DT)

When using DT, which is a binary branching approach, the model makes a comparison at each node based on a criterion, splits the data into two groups, and classifies the data points. This procedure continues until some previously defined stopping conditions are met. The main advantage of DT algorithm over KNN is that it can make predictions on dataset with complicated boundaries due to its branching capabilities. Therefore, it can create a chain of complex paths to classify the feature space into subsets with complex decision boundaries (Skiena, 2017). In this work, a multi-variate regression tree is implemented since the goal is to make prediction on multiple wave characteristics i.e., maximum, minimum, and average values for wave height and wave period at a given location.

For a given dataset with m data points, DT algorithm performs an exhaustive search to optimize the sum of squared errors for all the target variables, defined by

$$\sum_{i=1}^{q}\sum_{j=1}^{m}(y_i^{(j)} - \bar{y}_i)^2 \qquad (1)$$

In this equation, q is the number of target variables for the multi-output regression and $\bar{y}_i$ stands for the average value of target variable i for all the data points in the node. In this work, two stopping conditions are used. The maximum depth of the tree is 14 and the minimum number of samples required to split a node is six.

2.3 Support Vector Regression (SVR)

The SVR is the implementation of Support Vector Machine technique (Vapnik, 2013) for regression purposes. Two possible approaches can be adopted to develop a multi-output regression model. The first approach is to apply a single-output regression to each target variable and then develop a multi-output model using sing-output models (Yu, Yu, Tresp, & Kriegel, 2006). In this work, this approach is called Single-Target Support Vector Regression (STSVR). The second method is to chain together single-output models with each model using the input variables and the output of the previous model. In this work, this approach is referred to as Regressor Chain Support Vector Regression (RCSVR) (Zhang, Liu, Ding, & Shi, 2012).

The order of the target variables for RCSVR technique can affect the result and is typically decided either by a random selection or through training the model using various orders of target variables to find the best performing model (Spyromitros-Xioufis, Tsoumakas, Groves, & Vlahavas, 2012). It is worth mentioning that the drawback of STSVR is that it does not consider the possible dependency of target variables for model development and assumes target variables to be independent.

2.4 Neural Network (NN)

Capable of learning complex nonlinear systems, multi-layered NNs can greatly approximate any nonlinear mapping to any degree of accuracy. (Abbas Khosravi, 2011). They can be used for modeling, regression, and classification purposes. From one perspective, NNs are similar to biological nervous systems (Oludare Isaac Abiodun, 2018). In this work, a fully connected NN architecture is used for the purpose of multi-output regression modeling. The structure of the NN includes an input layer, an output layer, and some hidden layers. The best number of nodes at each layer was determined by trial and error, i.e. running the model with different number of nodes at each layer and looking for the best possible outputs for evaluation metrics.

2.5 Evaluation Metrics

Two evaluation metrics are used to assess the performance of the developed models, including Mean Squared Error (MSE) and R-squared score. The MSE for the multi-output regression models can be calculated using

$$MSE = \frac{1}{m}\sum_{i=1}^{q}\sum_{j=1}^{m}(y_i^{(j)} - \hat{y}_i^{(j)})^2 \tag{2}$$

where $\hat{y}_i^{(j)}$ stands for the predicted value of target variable i for data point j in a given dataset with m data points and q target variables (Brudnak, 2006). The R-squared score is defined as $R - Squared = 1 - aRSE$ with

$$aRSE = \frac{1}{q}\sum_{i=1}^{q}\frac{\sum_{j=1}^{m}(y_i^{(j)} - \hat{y}_i^{(j)})^2}{\sum_{j=1}^{m}(y_i^{(j)} - \bar{y}_i)^2} \tag{2}$$

where $\bar{y}_i$ represents the average for values of target variable i from the dataset (Aho, Ženko, Džeroski, & Elomaa, 2012).

3. Results and Discussion

Two types of multi-output regression models were developed in this work. The first type takes the latitude and longitude of a given location along with the month of the year and it then makes predictions for monthly maximum, monthly average, and monthly minimum values of the wave height in that location for that month. The second type of model takes similar inputs and makes predictions for monthly maximum, monthly average, and monthly minimum wave period for that site.

The required data for the analysis were collected from National Oceanic and Atmospheric Administration's National Data Buoy Center. The dataset includes the data from measurement buoys for the period 2010 to 2019 (10 years). Once the raw data was collected and before it was used for model development, it was wrangled and refined following the instructions provided by the World Meteorological Organization's Guidelines on the Calculation of Climate Normals (Organization, 2017). The data for a day of measurement were only counted if the missing observations were less than 2 hours during that day. Further, if the missing measurements for a month were more than 11 days or there were more than five consecutive days with missing measurements, the day was not taken into account. Moreover, and in this work, a year was only counted if the measurement data were available for at least seven months. Following these criteria, the final dataset included the data for 104 stations with 6,952 monthly data points. Figure1 shows the stations on the map and includes a bar chart representing the number of data points for each month.

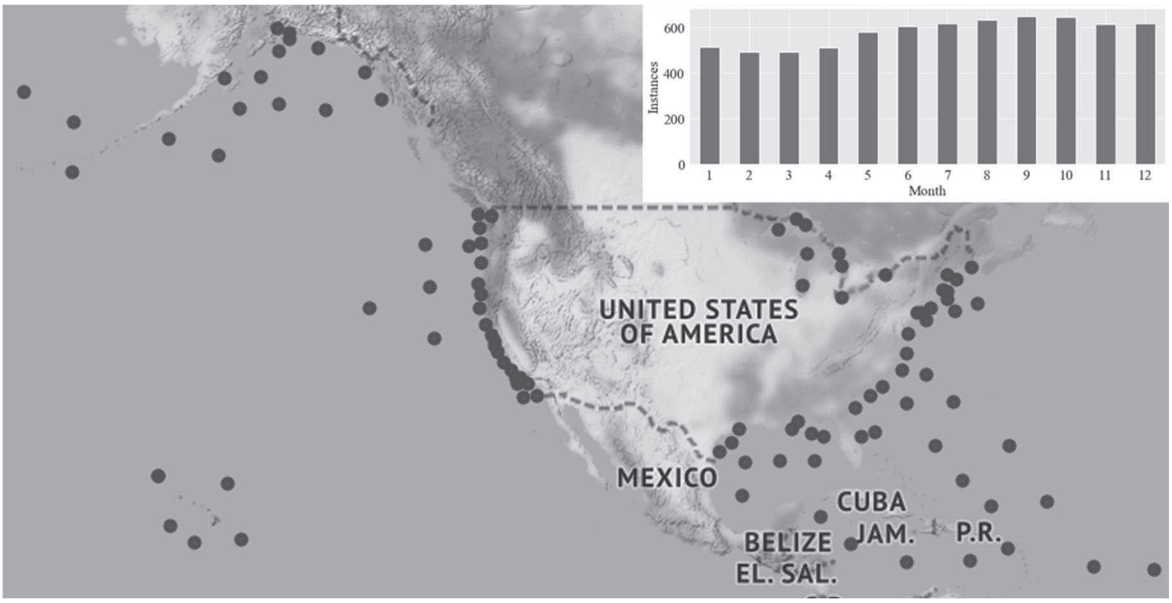


Fig. 1 The 104 stations with available data in the dataset and the number of data points for each month in the dataset

Two types of models were developed using KNN technique. In order to develop a KNN model, the best value for parameter K needs to be identified. In this work, a series of models were developed using KNN algorithm and the performance of each model was assessed using both MSE and R-squared score. Figures 2 and 3 show the results for models developed to predict both monthly wave height characteristics and monthly wave period characteristics. Based on these results, the k value of KNN model for wave height predicting model was set to five and the value for wave period predicting model was set to six.

To develop the NN models, a structure with four hidden layers, one input layer with three nodes, and one output layer with three nodes was adopted. The first three hidden layers had 100 nodes and the last hidden layer had 15 nodes. Figure 4(a) shows the general structure of the NN used for the modeling. The data was split into 80% training data, 10% validation data and 10% test data. The models were run with 4000 epochs and a batch size of 512. The Root Mean Square Error Propagation (RMSProp) was used as the optimizer and MSE was the loss function for the model. Figure 4(b) shows the loss function for 4,000 epochs.

The models were developed to make predictions on monthly wave height characteristics using all the techniques introduced in section 2. The results of the comparisons made between the measured values and the predicted values are shown in figure 5 for wave height values. Further, the results of the comparison between the measured and predicted values of monthly maximum wave periods, monthly average wave periods, and monthly minimum wave periods are shown in figure 6.

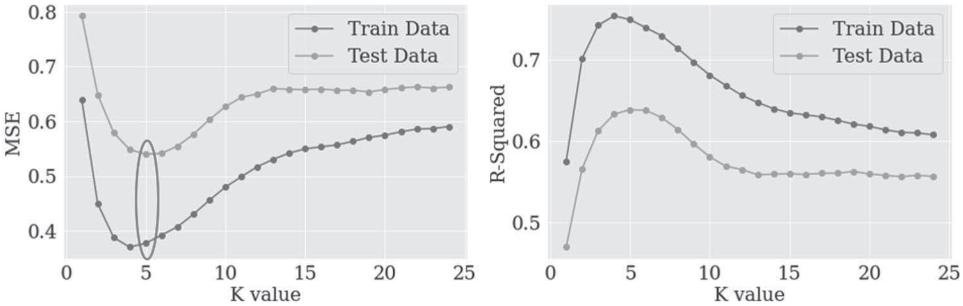


Fig. 2 The values of MSE and R-Squared for different K values (wave height model)

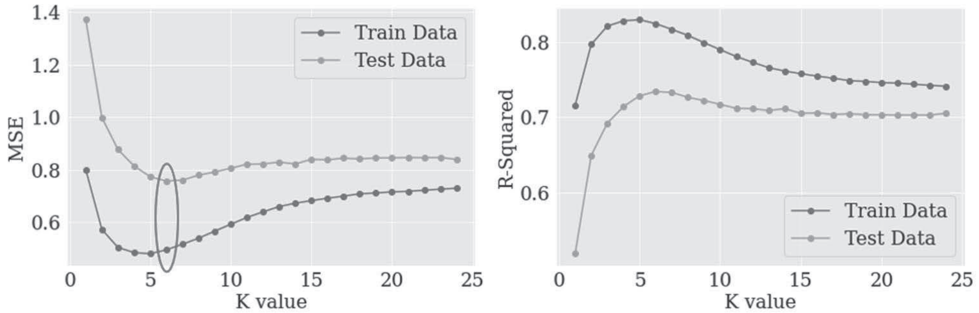
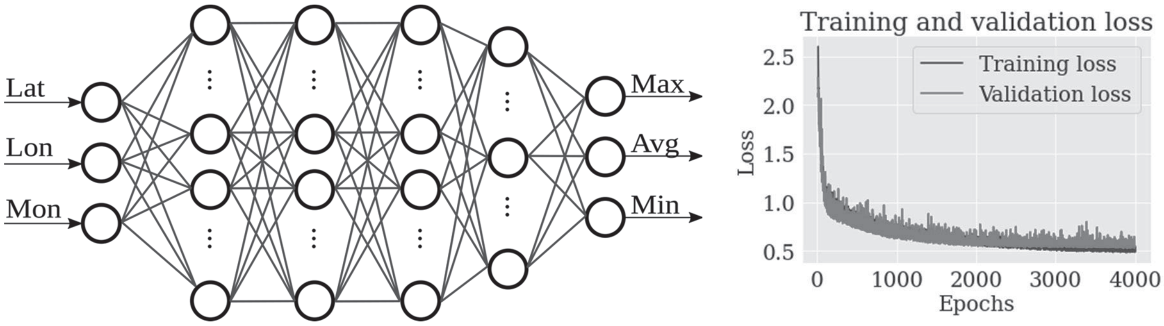
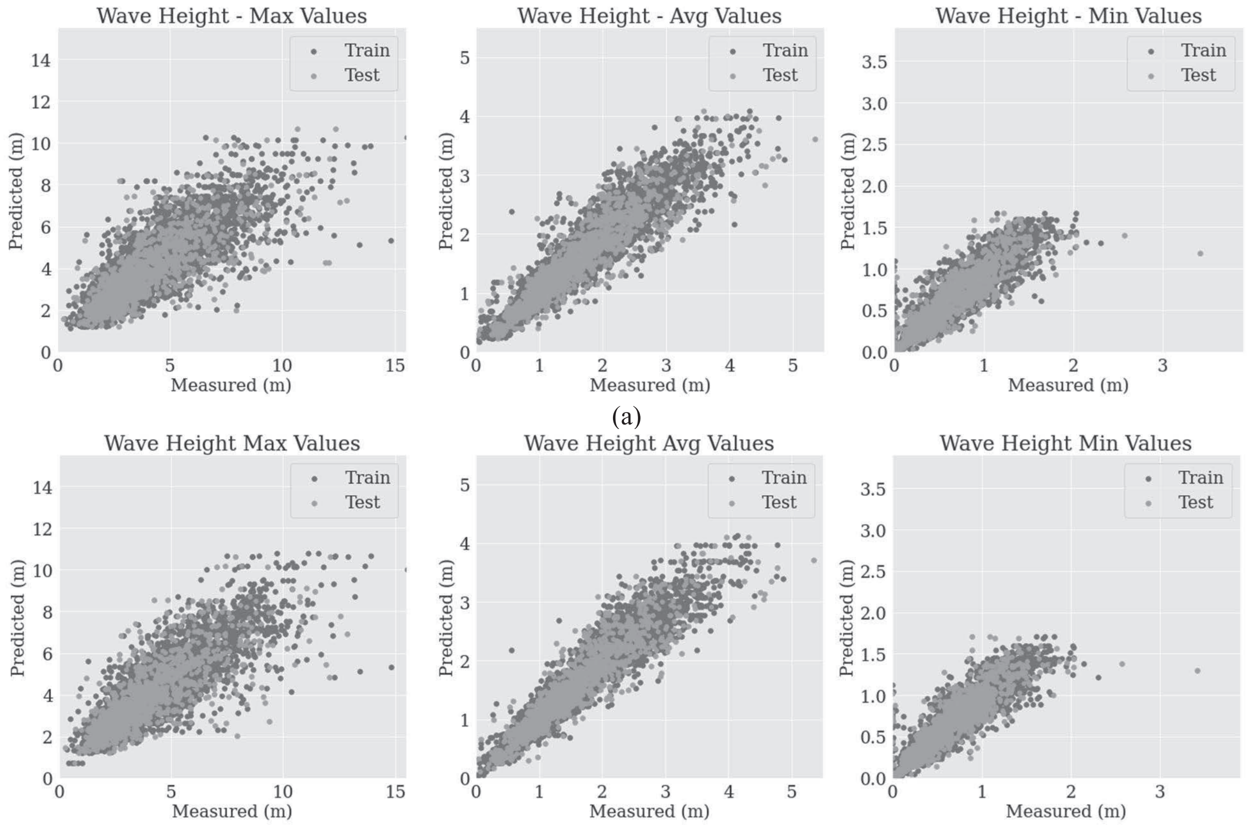


Fig. 3 The values of MSE and R-Squared for different K values (wave period model)



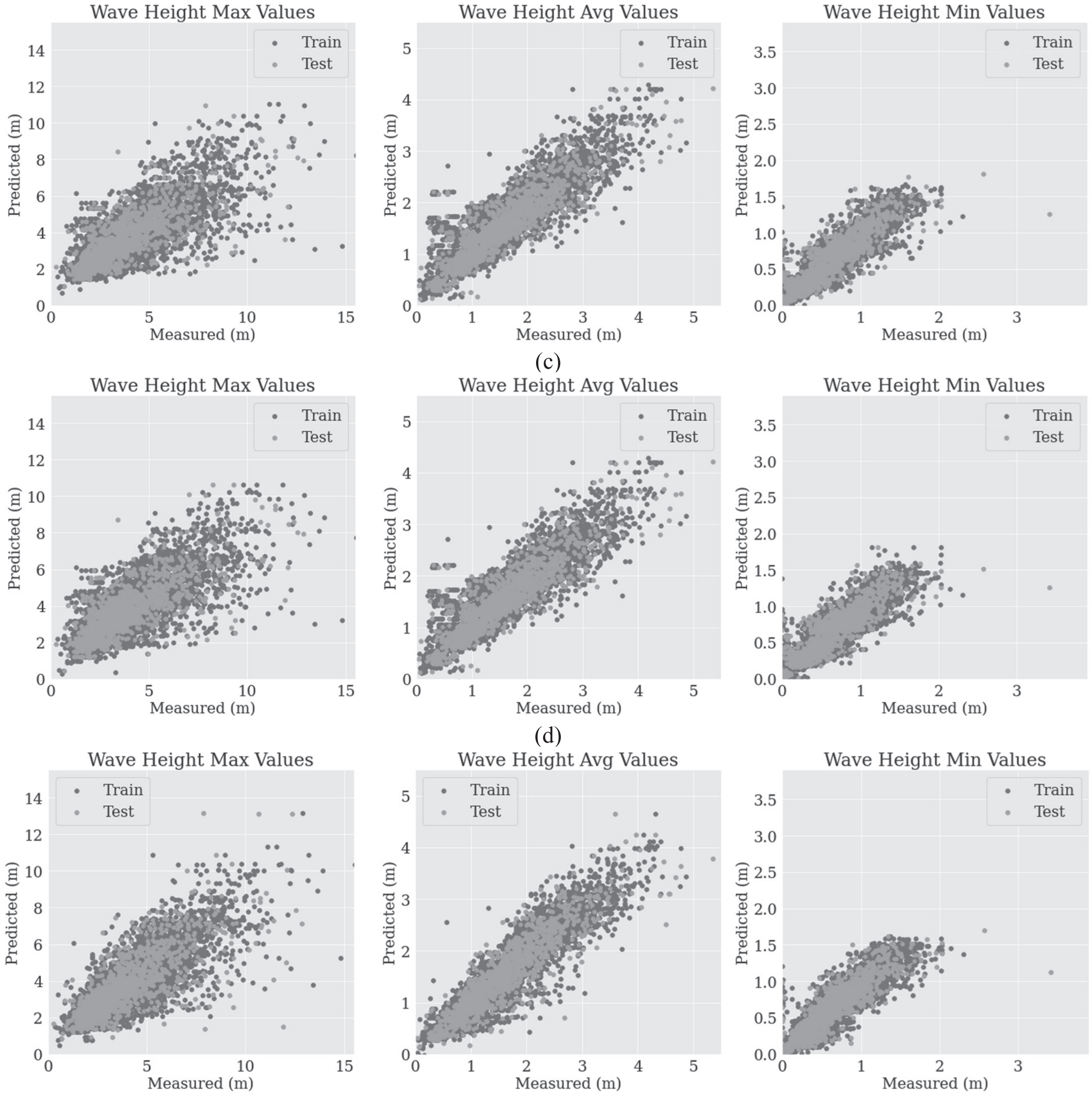
(a) (b)

Fig. 4 (a) The structure of NN, and (b) loss function for training and validation



(a)

(b)



(c)

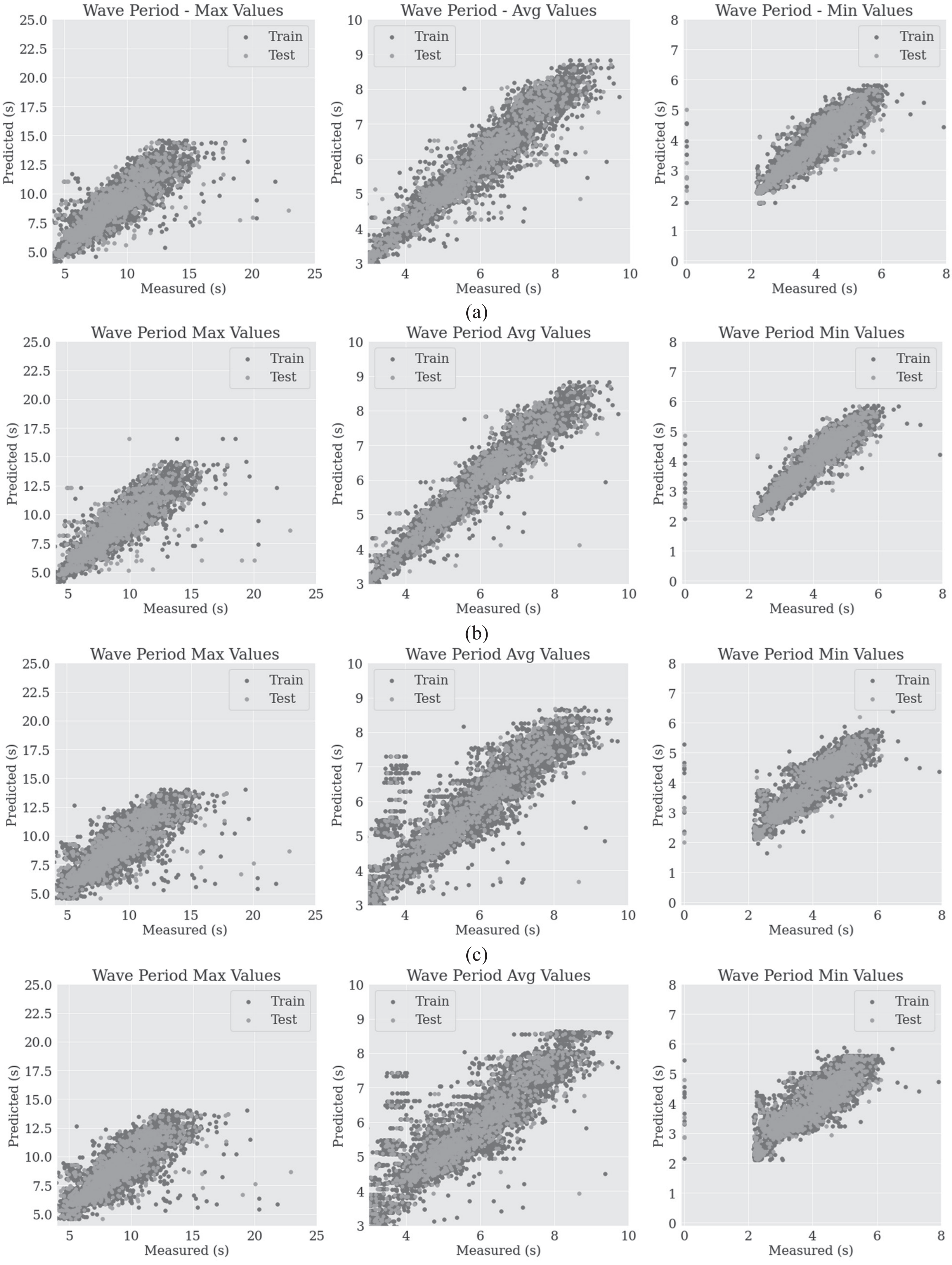
(d)

(e)

Fig. 5 Measured versus predicted values of wave height using (a) KNN model, (b) DT, (c) STSVR, (d) RCSVR, and (e) NN

Based on the results of the developed models, it can be seen that both types of models, i.e. wave height predicting models and wave period predicting models, are more effective in predicting average and minimum values. To have a better understanding of the performance of each model, the evaluation metrics are shown in figures 7 and 8. In these figures, MSEs and R-Squared scores are shown for all the models, separated by each target variable. This separation provides a better understanding of each model’s performance.

Based on the results and in general, models developed to predict wave period had a better performance compared to the models developed to predict wave height. For monthly wave height prediction, DT had the best performance in general. However, the KNN model also had a good performance when it came to monthly average and minimum wave height prediction. In terms of the wave period prediction and for monthly minimum and average values, the models developed using DT algorithm provided the lowest errors and highest scores. However, the models developed using KNN and NN could perform better in terms of predicting monthly maximum wave periods.

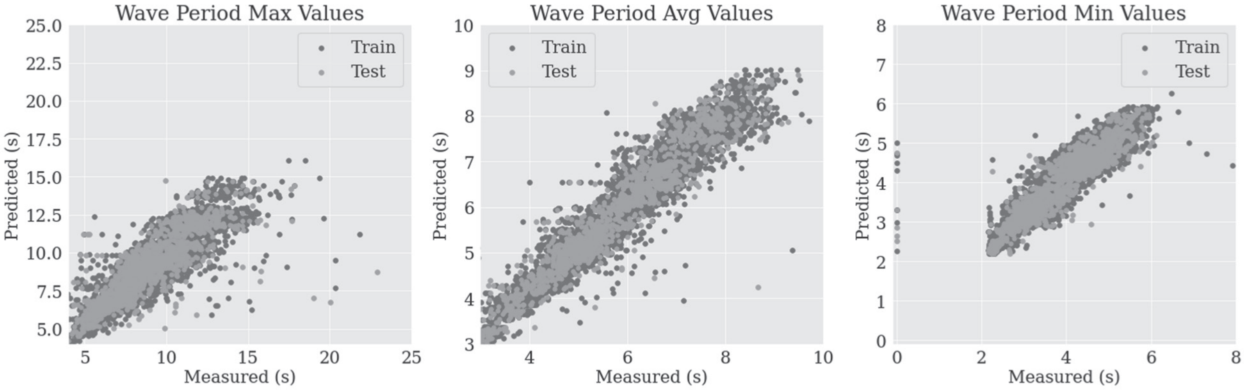

Wave Period - Max Values
Wave Period - Avg Values
Wave Period - Min Values
Wave Period Max Values
Wave Period Avg Values
Wave Period Min Values
Train
Test
Predicted (s)
Measured (s)
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
5
10
15
20
25
3
4
5
6
7
8
9
10
0
1
2

(a)

(b)

(c)

(d)



(e)

Fig. 6 Measured versus predicted values of wave periods using (a) KNN model, (b) DT, (c) STSVR, (d) RCSVR, and (e) NN

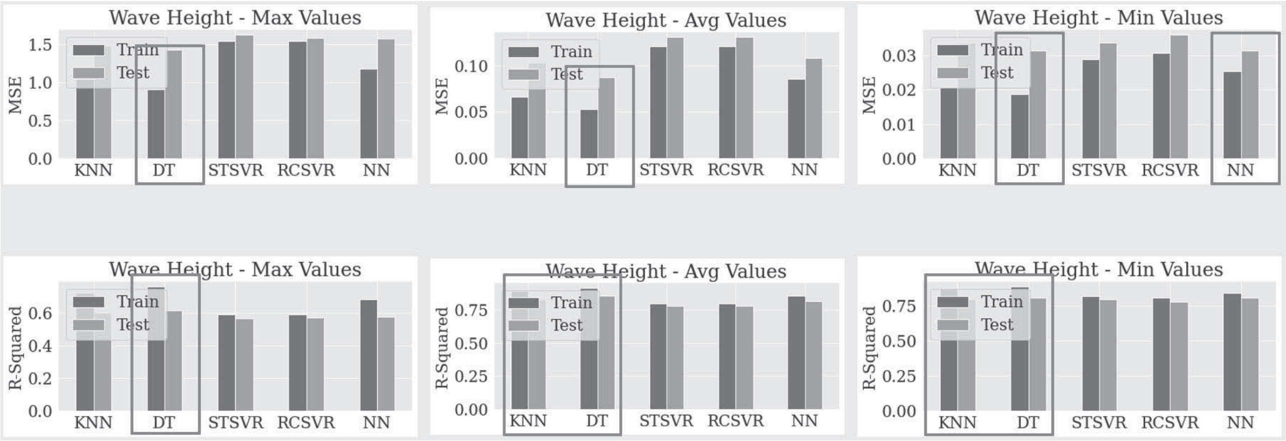


Fig. 7 Evaluation metrics MSE and R-Squared for the all wave height predicting models

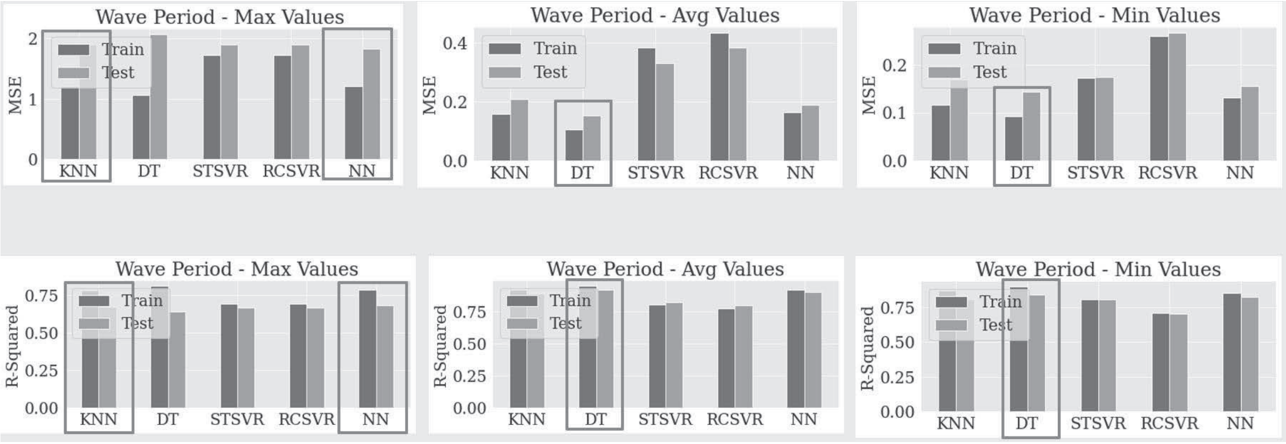


Fig. 8 Evaluation metrics MSE and R-Squared for the all wave period predicting models

4. Conclusion

Among the models implemented in this work, DT-, KNN- and NN-based models had the best performance in predicting all the features. The NN and KNN models outperformed other models in terms of predicting maximum values of wave periods. As it is expected, the models perform poorly in the areas with limited data as well as for the extreme conditions such as maximum wave height values. One observation was that the models developed using SVR algorithm did not perform well. One possible improvement for SVR-based modeling is to construct multi-output SVR models directly by accumulating target variables, and not transforming them into multiple single-output regression models.

Another possible approach to improve the accuracy of predictions is to cluster the coastal regions into similar regions and develop models for each cluster with similar features.

Models introduced here were much less accurate than more complex models (previously introduced in the literature), focused on a specific region. However, these models can be easily implemented and run in a few minutes on a typical computer, which makes them great candidates for developing a data-driven wave energy conversion system design tool.

References

Abbas Khosravi, S. N. (2011). Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances. *IEEE TRANSACTIONS ON NEURAL NETWORKS, 22*(9), 1341-1356.

Aho, T., Ženko, B., Džeroski, S., & Elomaa, T. (2012). Multi-target regression with rule ensembles. *The Journal of Machine Learning Research, 13*, 2367–2407.

Ali, M., & Prasad, R. (2019). Significant wave height forecasting via an extreme learning machine model integrated with improved complete ensemble empirical mode decomposition. *Renewable and Sustainable Energy Reviews, 104*, 281–295.

Berbić, J., Ocvirk, E., Carević, D., & Lončar, G. (2017). Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia, 59*, 331–349.

Brudnak, M. (2006). Vector-valued support vector regression. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, (pp. 1562–1569).

Cover, T. (1968). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory, 14*, 50–55.

Drew, B., Plummer, A. R., & Sahinkaya, M. N. (2009). A review of wave energy converter technology. *A review of wave energy converter technology*. Sage Publications Sage UK: London, England.

Karabulut, N., & Ozmen Koca, G. (2020). Wave height prediction with single input parameter by using regression methods. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, 1–18.

Kramer, O. (2013). K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors* (pp. 13–23). Springer.

Kumar, N. K., Savitha, R., & Al Mamun, A. (2018). Ocean wave characteristics prediction and its load estimation on marine structures: A transfer learning approach. *Marine Structures, 61*, 202–219.

Li, G., Weiss, G., Mueller, M., Townley, S., & Belmont, M. R. (2012). Wave energy converter control by wave prediction and dynamic programming. *Renewable Energy, 48*, 392–403.

Malekmohamadi, I., Bazargan-Lari, M. R., Kerachian, R., Nikoo, M. R., & Fallahnia, M. (2011). Evaluating the efficacy of SVMs, BNs, ANNs and ANFIS in wave height prediction. *Ocean Engineering, 38*, 487–497.

Mandal, S., & Prabaharan, N. (2010). Ocean wave prediction using numerical and neural network models.

N. Krishna Kumar, R. A. (2018). Ocean wave height prediction using ensemble of Extreme Learning Machine. *Neurocomputing, 277*, 12-20.

Nikoo, M. R., Kerachian, R., & Alizadeh, M. R. (2018). A fuzzy KNN-based model for significant wave height prediction in large lakes. *Oceanologia, 60*, 153–168.

Oludare Isaac Abiodun, A. J. (2018). State-of-the-art in artificial neural network applications: A survey. *Heliyon, 4*(11), 1-41.

Organization, W. M. (2017). WMO guidelines on the calculation of climate normals. *WMO guidelines on the calculation of climate normals*. World Meteorological Organization Geneva, Switzerland.

Pirhooshyaran, M., & Snyder, L. V. (2020). Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence-to-sequence networks. *Ocean Engineering, 207*, 107424.

Ribeiro, G. T., Mariani, V. C., & dos Santos Coelho, L. (2019). Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting. *Engineering Applications of Artificial Intelligence, 82*, 272–281.

Savitha, R., Al Mamun, A., & others. (2017). Regional ocean wave height prediction using sequential learning neural networks. *Ocean Engineering, 129*, 605–612.

Skiena, S. S. (2017). *The data science design manual.* Springer.

Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing, 251*, 26–34.

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., & Vlahavas, I. (2012). Multi-label classification methods for multi-target regression. *arXiv preprint arXiv:1211.6581*, 1159–1168.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

Yu, S., Yu, K., Tresp, V., & Kriegel, H.-P. (2006). Multi-output regularized feature projection. *IEEE Transactions on Knowledge and Data Engineering, 18*, 1600–1613.

Zamani, A., Solomatine, D., Azimian, A., & Heemink, A. (2008). Learning from data for wind–wave forecasting. *Ocean engineering, 35*, 953–962.

Zhang, W., Liu, X., Ding, Y., & Shi, D. (2012). Multi-output LS-SVR machine in extended feature space. *2012 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA) Proceedings*, (pp. 130–134).