International Journal of Engineering & Technology, 7 (4.41) (2018) 91-95

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Comparative Analysis of Machine Learning Models for Prediction of Wave Heights in Large Waterbodies

Priyanka Sinha[1], Shweta Vincent[1], Om Prakash Kumar[1*]

[1]Manipal Institute of Technology, MAHE, Karnataka INDIA
**Corresponding author E-mail:omprakash.kumar@manipal.edu*

Abstract

This paper presents a study of the various machine learning algorithms viz. Linear Regression, Logistic Regression, Support Vector Machine, Support Vector Regression and Extreme Machine Learning for the prediction of wave heights using data obtained from ocean buoys. The data from the ocean buoy number 62081 off the coast of Ireland in Europe has been chosen for study. It is found that the parameter of wind speed affects wave heights the most in comparison to other parameters. It is also observed that Extreme Learning Machine outperforms Support Vector Regression when classifying the data points as high tide or low tide. The MSE and CC parameters prove the suitability of Extreme Machine Learning over all the other algorithms discussed in this paper for the accurate prediction of wave heights.

Keywords: *Linear Regression, Logitstic Regression, Support Vector Machine, Support Vector Regression, Extreme Learning Machine*

1. Introduction

The prediction of wave heights is instrumental in planning of offshore activities in the case of an emergency such as a tornado or tsunami. Complex numerical algorithms are used for the prediction of wave heights in seas and oceans. In the past, several wave models have been numerically developed using the energy balance equation. Nonlinear wave interaction posed the greatest challenge in terms of analysis [1].

The third generation wave models have utilized the components of the source function without any prior restrictions on the spectral shape. There is a huge scope for improvement of models for better representation of complex physical processes leading to waves generated by winds. With the advent of Machine Learning and Neural Networks, several algorithms have been devised in order to make faster predictions and they are also computationally efficient. This article aims at optimizing the use of wave parameters for the prediction of wave heights in water bodies in order to further assist offshore activities. There are no assumptions to be made, or boundary conditions to be considered in wave height prediction [3] using neural networks as opposed to using complex numerical models. The neural network does not recognize the physical phenomena. It establishes the relationships between inputs and outputs based on learning processes.

Soft computing techniques namely, SVM, BNs, ANNs and ANFIS have been applied in order to determine wind height (WH) using existing wind speed data using a buoy at Lake Superior [5]. ANN, FIS and ANFIS have also been compared for wave height prediction at Lake Ontario. It was demonstrated that ANFIS yielded better results in comparison to ANN and FIS.

Both the algorithms of BNs and ANFIS are used when the parameters of probability and confidence are both important for wave height prediction. Also, BNs and SVM are capable of handling uncertainties in the input-output pattern of variables under consideration. BNs are applied in problems when the exact value of one (or more) input variables is not available, as opposed to, SVM, ANN and ANFIS [4, 5].

The first section of this article describes the theoretical background behind our proposed system and presents the algorithms used for the project. The second section presents in the detail the experiments conducted in order to predict wave heights using the machine learning algorithms. The third section is dedicated to discussion of the results achieved and a discussion on the performance parameters. The final section concludes the article and presents the scope for future work in this area.

2. Theoretical Background

2.1. Spectral Energy Balance Equation

The Spectral Energy Balance equation which is represented in Equation (1) forms the basis for the numerical model for wave height prediction.

It is represented as:

$$\frac{\partial E(f,\theta,t)}{\partial t} = S = S_{in} + S_{nl} + S_{ds} \tag{1}$$

where, $\frac{\partial E(f,\theta,t)}{\partial t}$ is the spectrum of the wave which depends on the frequency f and the direction of propagation, θ . The net source function is represented by S and it depends on the factor S_{in} , which are the external wave making factors such as local wind and local current, S_{nl} is the non-linear energy conduction by wave-wave interactions and S_{ds} the dissipation related to wave-disperse processes and its reaction with turbulence of the water layer on the surface.

2.2. Linear Regression

Linear Regression is an approach for arriving upon a relation between a scalar dependent variable *y* and one (simple linear regression) or more (multiple linear regression) independent variables denoted as *X*. If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. The least squares error approach is used to fit a line to a set of data point in Linear Regression. The general hypothesis function for Linear Regression is represented as shown in Equation (2).

$$\hat{Y} = \theta_0 + \theta_1 X \tag{2}$$

Here $\hat{Y}$ is the predicted value, θ_0, θ_1 are the weights of the line and X is the set of input variables. The accuracy of the line fit can be estimated using a Cost function represented in Equation (3).

$$J(\theta_0, \theta_1) = \frac{\sum_{i=1}^{m} (\hat{y}_i - y_i)^2}{2m} \tag{3}$$

where, $J(\theta_0, \theta_1)$ is the cost function, $\hat{y}_i - y_i$ is the difference between the predicted and actual values of *y*. *m* is the number of data points available for fitting the line.

2.3. Logistic Regression

Logistic Regression is similar to Linear Regression in terms of its approach. The difference lies in the fact that curves other than straight lines can be fit using Logistic Regression.

The hypothesis function for Logistic Regression is given in Equation (4).

$$h_\theta(x) = g(\theta^T X) \tag{4}$$

where, if $z = \theta^T X$, then $g(z) = \frac{1}{1+e^{-z}}$. The function $g(\theta^T X)$ represents the Sigmoid function such that the conditions of Equation (5) get satisfied.

$$h_\theta(x) > 1;\ y = 1\ \ else\ \ h_\theta(x) < 1;\ y = 0 \tag{5}$$

Similar to Linear Regression, even in Logistic Regression, a cost function is computed and the Gradient Descent algorithm is applied over the cost function to find the least cost fitting curve to a particular data set.

In our article, both the algorithms of Linear and Logistic Regression have been explored and the detailed discussions of the results obtained are presented in Section 3 of this article.

2.4. Support Vector Machine

A Support Vector Machine (SVM) algorithm is used as a supervised learning model for classification and analysis of regression models. When a set of training examples is given to this algorithm, it performs binary linear classification on the dataset. SVM is also capable of performing non-linear classification using a kernel trick by mapping inputs into higher dimensional feature spaces. The Support Vector Clustering algorithm utilizes the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data. It is widely used for industrial applications.

2.5. Support Vector Regression

Given a training data set $\{(x_1, y_1), ...(x_n, y_n)$ where, the set *X* denotes the inputs, then the goal of Support Vector Regression (SVR) is to find a function *f(x)* such that it has utmost ε deviation from the actually obtained targets y_i for all the training data and is also as flat as possible [7].

The performance of SVR is determined by the shape of the kernel function and other parameters which could represent the noise in the training data. Advanced methods using Bayesian networks could be employed to determine the noise parameters in the training dataset.

2.6. Extreme Learning Machine

Extreme Learning Machine (ELM) is a branch of ANN which uses feed forward networks for the processes of classification and regression and other complex machine learning functions [2]. The hidden nodes of these networks could be assigned weights on a random basis and may never be updated or could be inherited from their ancestors.

According to their creators, these models are able to produce good generalization performance and learn thousands of times faster than networks trained using backpropagation. In literature, it also shows that these models can outperform support vector machines (SVM) [8].

The upcoming section of this article presents the experimentation carried out to predict the wave heights of the ocean, off the coast of Ireland, using data from ocean buoys present in the area.

3. Ocean Wave Height Prediction

EMODnet is a data portal for Europe which provides aspiring researchers with data related to oceanic activities around Europe. This data is gathered by placing ocean buoys *in-situ* which collect information about wind temperature, atmospheric pressure, wind speed etc. which are crucial parameters in predicting the height of the waves.

One such ocean buoy number 62081 as shown in Figure 1 has been used to tap data related to the ocean off the coast of Ireland. The data collected by these buoys is in real time and is available in NETcdf file format. The NETcdf library of MATLAB has been made use of for the analysis of this data.

Fig 1: Ocean buoy 62081 off the coast of Ireland used for data extraction

The list of parameters which are obtained by the ocean buoy are listed in Table 1.

Table 1: Ocean buoy Parameters

S. No.	Parameter Name
1	Wave Height
2	Air Temperature
3	Wind Speed
4	Sea Temperature
5	Atmospheric Pressure
6	Water Conductivity
7	Wind Direction
8	Salinity

The dataset was formed with the varied input parameters as X and ocean wave height as the output parameter. For linear regression, the three datasets which have been used are Wind Speed vs. Wave Height, Air Temperature vs. Wave Height and Sea Temperature vs. Wave Height.

For logistic regression, a dataset of Wind speed, Atmospheric Pressure and Wave Height was used. The wave heights above 8 meter were labeled as 1 and wave heights below 8 meter were labeled zero indicating 1 as the high tide and 0 as the low tide. For Support Vector Regression and Extreme Learning Machine, all the four parameters viz. Wind speed, Atmospheric Pressure, Air Temperature and Sea Temperature were loaded as the inputs along with the targets. This dataset was named as the training data set. The trained model was further tested on the test data. For Extreme Learning Machine, all the data was normalized.

Figure 2 below depicts a flow chart which gives the general flow followed in this project for the prediction of ocean wave heights.

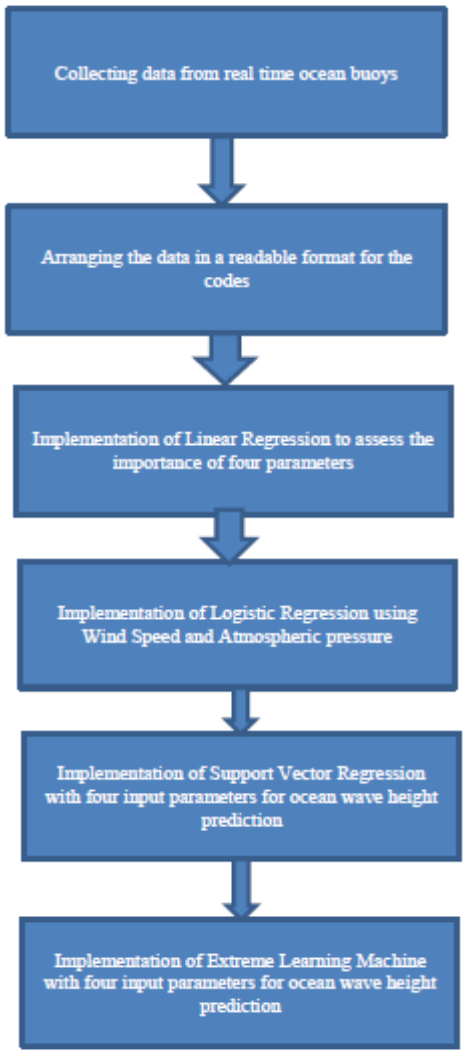

Fig 2: Flowchart depicting the project flow

4. Results and Discussion

This section of the article presents the results of the predictions and correlations obtained. The Linear Regression algorithm was first implemented on the data set by using the individual parameters listed in Table 1 as inputs *X* versus the Wave height. The importance of the parameters was assessed by looking at performance metrics of Root Mean Square Error (RMSE) and Correlation Coefficient (CC).

The Logistic Regression Algorithm was also implemented to perform a generalized prediction of 'High vs. Low' tide using two parameters namely Atmospheric Pressure and Wind Speed.

To perform an Ocean Wave Height Prediction using all the parameters, the algorithms of Support Vector Regression and Extreme Learning Machine were used.

4.1. Results of Linear Regression

The importance of the four parameters namely Air Temperature, Sea Temperature, Wind Speed and Atmospheric Pressure in wave height detection was evaluated using Linear Regression. Predictions were performed using all the four parameters separately. To assess the performance, parameters of Root mean square error (RMSE), Mean square error (MSE), R squared and Coefficient of Correlation (CC) were used. Lower the MSE values, better the performance. Higher the correlation coefficient, better the performance. Figure 3 illustrates the graph obtained plotting Wave height vs. Air Temperature. The conclusion drawn from Figure 3 is that there is very low correlation between Wave Height and Air Temperature. Figure 4 describes the prediction of Wave Height using data of Air Temperature. As is clear from the figure, Air temperature has a very low correlation with wave height.

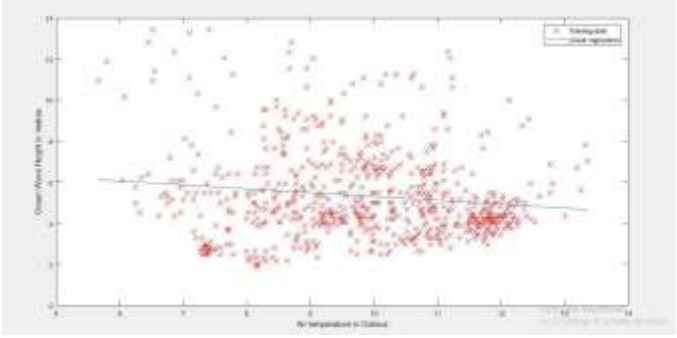

Fig 3: Linear Regression model of Wave Height vs. Air Temperature

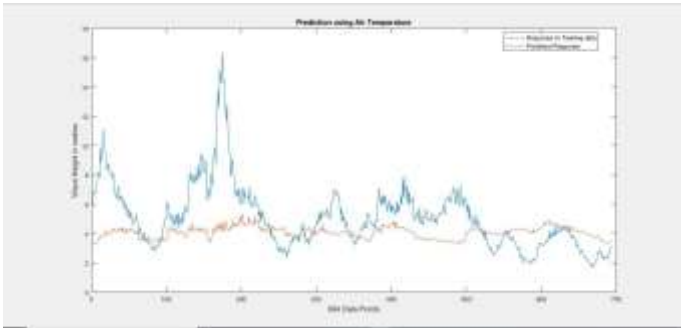

Fig 4: Wave height prediction using Air Temperature

Similar to the results obtained in Figures 3 and 4, Linear regression and predictions were performed for Wave Height vs. Atmospheric Pressure (Figures 5 and 6), Wave Height vs. Sea Temperature (Figures 7 and 8) and Wave Height vs. Wind Speed (Figures 9 and 10). As is clear from Figure 5 and 6, there is not much correlation between Wave Height and Atmospheric Pressure.

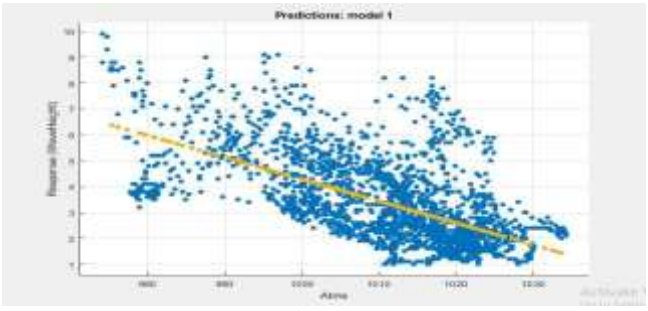

Fig 5: Linear Regression model of Wave Height vs. Atmospheric Pressure

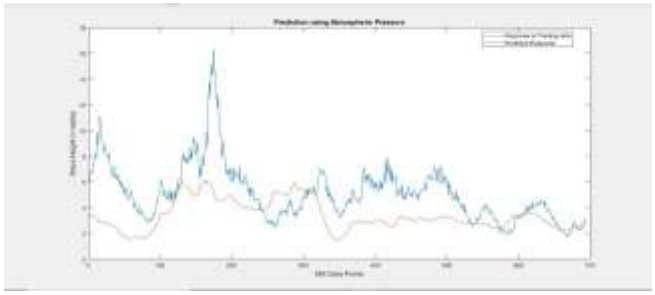

Fig 6: Wave height prediction using Atmospheric Pressure

In the case of Wave Height vs. Sea Temperature a negative correlation is obtained as depicted in Figure 7. Figure 8 depicts a bad prediction of wave height using only sea temperature.

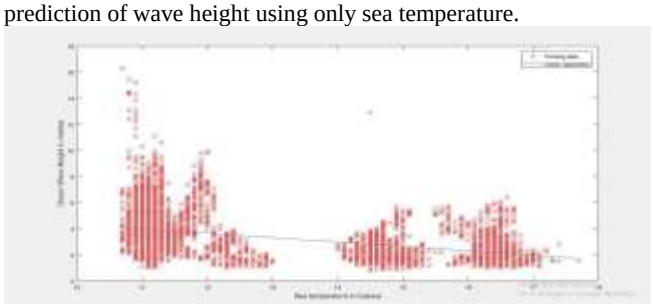

Fig 7: Linear Regression model of Wave Height vs. Sea Temperature

Finally the plot of Figure 9 gives a positive correlation between Wave Height and Wind speed. Figure 10 also illustrates a good prediction of wave height using wind speed only.
Therefore, it was concluded that of all parameters, Wind Speed has the highest correlation with Wave heights. Table 2 describes the comparative analysis of the aforementioned results.

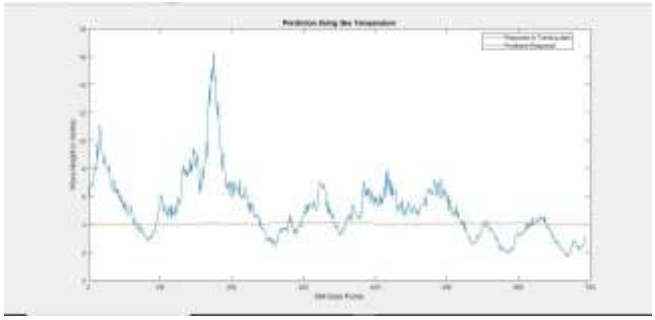

Fig 8: Wave height prediction using Sea Temperature

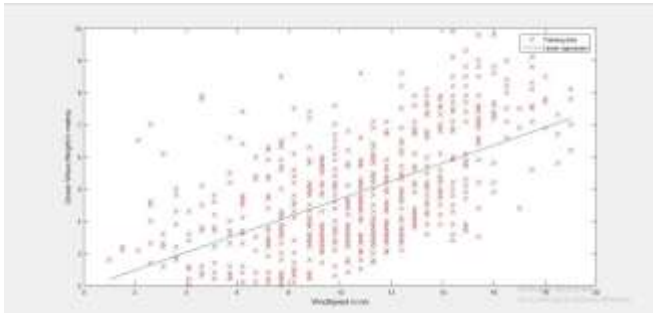

Fig 9: Linear Regression model of Wave Height vs. Atmospheric Pressure

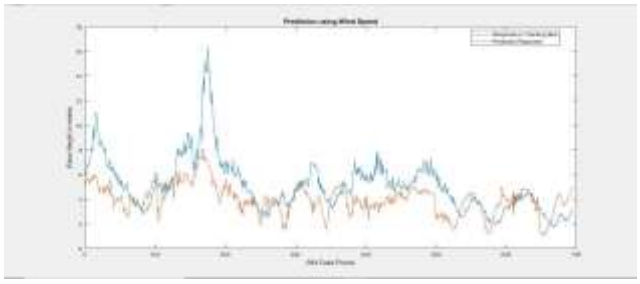

Fig 10: Wave height prediction using Wind Speed

Table 2: Comparative Analysis of parameters affecting Wave Heights

Parameter	RMSE	R squared	MSE	CC
Wind Speed	1.2366	0.50	1.5292	0.6354
Atmospheric Pressure	1.3413	0.41	1.7911	0.3977
Air Temperature	1.6063	0.16	2.5801	0.1711
Sea Temperature	1.5763	0.19	2.4847	0.0388

4.2. Results of Logistic Regression and Support Vector Machine

A general classification boundary was generated for the data points using two parameters namely Atmospheric Pressure and Wind Speed. A wave height of below 8 meters was classified as '0' and the rest of the data points were classified as '1'. Figure 11 depicts the classification results obtained by pure Logistic regression to classify the data wave heights as high tide or low tide. Figure 12 depicts the classification results of the same dataset using Support Vector Machine. This shows a higher accuracy in classification in comparison to Logistic Regression. Figure 13 depicts the classification of the data points which was performed using Gaussian Kernel Support Vector Machine. Of all three methods, the latter showed the best results in terms of accuracy of classification.

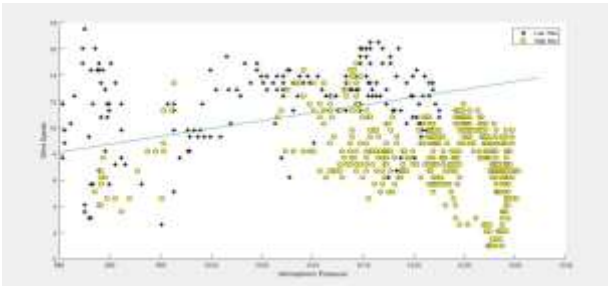

Fig 11: Classification of High Tide and Low Tide using Logistic Regression

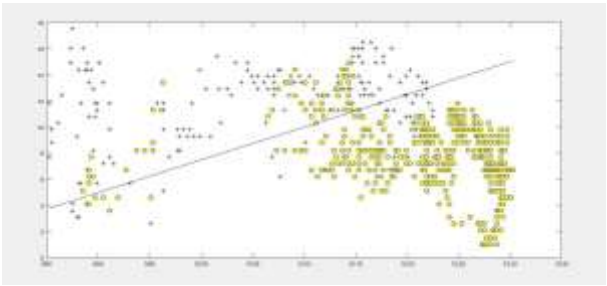

Fig 12: Classification of High Tide and Low Tide using Linear Support Vector Machine

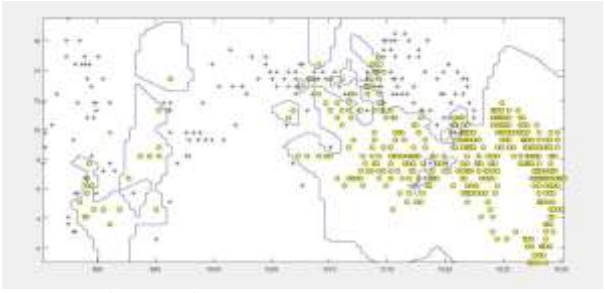

Fig 13: Classification of High Tide and Low Tide using Gaussian Kernel Support Vector Machine

Table 3 illustrates the consolidated results of all three algorithms.

4.3. Results of Support Vector Regression

The performance of the Support Vector Regression (SVR) was analyzed using the parameters of Mean Square Error (MSE), Co-

efficient of Correlation (CC) and Epsilon Loss Insensitive Value (EL) on the data collected from the buoy. A higher value of EL indicates a higher tolerance towards errors.

Table 3: Accuracy of classification of High tide and Low tide data

Algorithm	Classification Accuracy
Logistic Regression	85%
Support Vector Machine	84.8%
Gaussian Kernel Support Vector Machine	88%

It was observed that by increasing the epsilon value the accuracy of the classification results increased considerably. This is illustrated in Figure 14 and the values of the corresponding parameters are tabulated in Table 4.

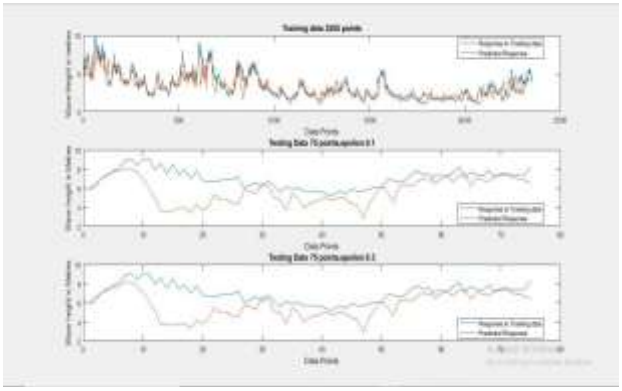

Fig 14: Support Vector Regression plot: Predicted vs. Actual data

Table 4: Parameters of Support Vector Regression

Epsilon Value	MSE	CC	EL
0.1	0.6564	0.4291	0.3228
0.3	0.6477	0.4358	0.4654

4.4. Results of Extreme Learning Machine

The Extreme Learning Machine algorithm was applied to the problem of classification of high tide and low tide. An ELM of 40 hidden neurons was used with a Sigmoidal activation function. The MSE obtained by this was 0.1321 with a CC of 0.6. Figure 15 showcases the result of the predicted output. Table 5 tabulates the comparative results of SVR and ELM.

Table 5: Comparative Results of SVR and ELM classification

Algorithm	MSE	CC
SVR	0.6477	0.4358
ELM	0.1321	0.6690

The low value of MSE and high value of CC lead to the conclusion that, ELM outperforms all the other algorithms viz. Linear Regression, Logistic Regression, Support Vector Machine and Support Vector Regression. Table 6 illustrates the comparative analysis of the number of false positives and false negatives in the predictions made by ELM and SVR. In this table also, ELM outperforms SVR.

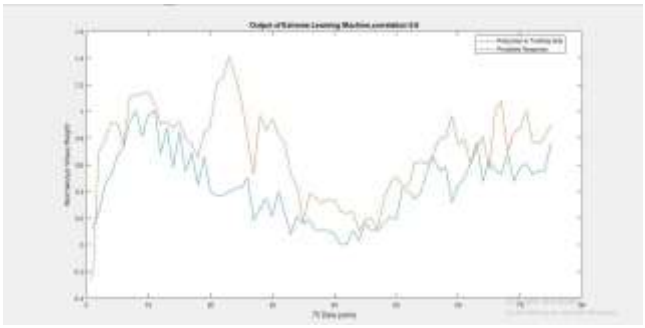

Fig 15: Extreme Learning Machine plot: Predicted vs. Actual data

Table 6: Comparative Results of False Positives and False Negatives for SVR and ELM classification

Algorithm	False Positives	False Negatives
SVR	64	11
ELM	71	4

5. Conclusion and Future Work

This paper presented a study of the various machine learning algorithms which can be used for the prediction of wave heights using data obtained from ocean buoys. The data from the ocean buoy number 62081 off the coast of Ireland in Europe was chosen for study. This buoy provided data for the parameters tabulated in Table 1. The individual parameters were plotted against wave height using a Linear Regression model and it was observed that the parameter of Wind Speed affected the wave height the most when compared to the other parameters. An analysis of the results obtained from Logistic Regression, Support Vector Machine and Gaussian Kernel Support Vector Machine was performed to classify the wave height data as either high tide or low tide. It was observed that, the Gaussian Kernel Support Vector Machine provided the highest accuracy of 88%.

The parameter of Epsilon Loss was explored for the algorithm of Support Vector Regression and it was observed that the accuracy of the classification increased by increasing the value of epsilon.

Finally, the Support Vector Regression algorithm was compared with the Extreme Learning Machine algorithm and it was observed that the latter outperformed the former in terms of MSE, CC and False positive and negative numbers.

This work can be further extended to perform a generalized prediction using Reinforcement Learning over all water bodies. Further, the next aim of this work should be to attain more accuracy in prediction. The world is facing a lot of problems because of which the water levels in the seas and oceans are rising. If the rise in these levels could be predicted accurately well beforehand, then a number of catastrophes could be prevented or mitigated.

References

[1] N. Kumar, R. Savitha and A. Al Mamun, "Ocean wave height prediction using ensemble of Extreme Learning Machine", Journal of Neurocomputing, vol. 277, pp. 12-20, 2018.

[2] G. Huang, Q. Zhu and C. Siew, "Extreme learning machine: Theory and applications", Journal of Neurocomputing, vol. 70, no. 1-3, pp. 489-501, 2006.

[3] S. Londhe and V. Panchang, "One-Day Wave Forecasts Based on Artificial Neural Networks", Journal of Atmospheric and Oceanic Technology, vol. 23, no. 11, pp. 1593-1603, 2006.

[4] M. Kazeminezhad, A. Etemad-Shahidi and S. Mousavi, "Application of fuzzy inference system in the prediction of wave parameters", Ocean Engineering, vol. 32, no. 14-15, pp. 1709-1725, 2005.

[5] I. Malekmohamadi, M. Bazargan-Lari, R. Kerachian, M. Nikoo and M. Fallahnia, "Evaluating the efficacy of SVMs, BNs, ANNs and ANFIS in wave height prediction", Ocean Engineering, vol. 38, no. 2-3, pp. 487-497, 2011.

[6] A. Durán-Rosal, C. Hervás-Martínez, A. Tallón-Ballesteros, A. Martínez-Estudillo and S. Salcedo-Sanz, "Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks", Ocean Engineering, vol. 117, pp. 292-301, 2016.

[7] S. Salcedo-Sanz, J. Nieto Borge, L. Carro-Calvo, L. Cuadra, K. Hessner and E. Alexandre, "Significant wave height estimation using SVR algorithms and shadowing information from simulated and real measured X-band radar images of the sea surface", Ocean Engineering, vol. 101, pp. 244-253, 2015.

[8] A. Smola and B. Schölkopf, "A tutorial on support vector regression", Statistics and Computing, vol. 14, no. 3, pp. 199-222, 2004.