INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

OPEN ACCESS

Received: 19-04-2023

Accepted: 03-07-2023

Published: 29-08-2023

Citation: Alqahtani AM, Elbisy MS, Osra FA (2023) Comparing Between Support Vector Machine and Gradient Boosted Trees Models for Prediction of Wave Overtopping at Coastal Structures with Composite Slope. Indian Journal of Science and Technology 16(33): 2580-2588. https://doi.org/10.17485/IJST/v16i33.919

* **Corresponding author**.

s44280359@st.uqu.edu.sa

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Comparing Between Support Vector Machine and Gradient Boosted Trees Models for Prediction of Wave Overtopping at Coastal Structures with Composite Slope

Ammar M Alqahtani[1]*, Moussa S Elbisy[1], Faisal A Osra[1]

1 Civil Engineering Dept, College of Engineering and Islamic Architecture, Umm Al_Qura University, Makkah, Saudi Arabia

Abstract

Background/Objectives: Predicting wave overtopping at coastal structures is a critical task in coastal engineering. The use of machine learning models can help predict wave overtopping with higher accuracy and efficiency. **Methods**: In this study, the accuracy of support vector machine (SVM) and gradient-boosted tree (GBT) approaches for predicting wave overtopping discharge of coastal structures with composite slopes ”without a berm” was evaluated and compared. The newly developed EurOtop database was used for this study. **Findings**: The results showed that the GBT technique provided more accurate predictions than the SVM technique. Notably, the results demonstrated that the GBT model significantly decreased the overall error and accurately estimated the wave-overtopping discharge. Compared to the SVM model, the scatter index calculated using the GBT (0.57) is lower than that of the SVM (o.94). In terms of R-index, the GBT (0.97) was superior to the SVM (0.92) model. **Novelty**: This study provides an effective approach for predicting wave overtopping discharge in the design and management of coastal structures.

Keywords: Prediction; Wave Overtopping; Machine Learning Techniques; Coastal Structures; Safety

1 Introduction

Wave overtopping prediction is the process of estimating the quantity of water that may spill over a coastal structure during a storm or high wave event. This is an important endeavor for ensuring the safety and functionality of coastal structures, such as seawalls, breakwaters, and other types of maritime infrastructure. The traditional deterministic models based on physical principles and mathematical models have limitations in handling the uncertainties and complexities associated with the wave overtopping prediction. This has led to the application of soft computing techniques in this area. These techniques have also reduced the computational time associated with the prediction process. Therefore, the use of machine learning techniques for wave overtopping prediction is expected to continue to grow in the future.

Artificial Neural networks (ANN) have become increasingly popular in the field of wave overtopping prediction due to their ability to learn complex relationships and patterns in data. Several studies have employed neural networks for this purpose, including Pullen et al.[1], Verhaeghe et al.[2], Molines and Medina[3] and Formentin et al.[4]. These studies have shown that neural networks can be effective in predicting wave overtopping, with some achieving high levels of accuracy in their predictions. One advantage of neural networks is their ability to handle large amounts of data and learn complex relationships between input and output variables, which can be particularly useful in the context of wave overtopping prediction where many factors can influence the outcome. Overall, neural networks have become an important tool in wave overtopping prediction, and their use is likely to continue to grow as more data becomes available and computing power continues to increase.

Furthermore, den Bieman et al.[5] recently demonstrated that the XGBoost method[6] could be effectively used as an alternative to ANN models. Habib et al.[7] reviewed the use of ML to predict wave overtopping discharge and overtopping parameters. They highlighted the important limitations of the methods and identified future research needs that may serve as an impetus for the further development of these machine learning algorithms for wave-overtopping, particularly in applications characterized by complex geometrical configurations.

Elbisy[8] conducted a study on the estimation of wave-overtopping discharge at coastal structures with a straight slope using different machine learning algorithms. Specifically, they used multilayer perceptron (MPNN), cascade correlation neural network (CCNN), and GRNN, as well as support vector machines (SVMs) for this purpose. Overall, the study suggests that the GRNN model, can be a useful tool for predicting wave-overtopping discharge at coastal structures with a straight slope without a berm. The study's findings could have practical applications in the field of coastal engineering, such as in designing and optimizing coastal structures to withstand wave-overtopping. Alshahri and Elbisy[8] employed artificial neural network-based (ANN) approaches with different algorithms, such as MPNN, and GRNN, and SVM for estimating the wave overtopping discharge at rubble mound structures featuring a straight slope. They found that the GRNN produced exceptionally precise results.

The goal of this work is to provide coastal designers with a robust and accurate machine learning model able to represent wave overtopping discharges for a wide range of coastal structure types with composite slopes under a variety of wave conditions. This study evaluates the accuracy of machine learning models utilizing the support vector machine and gradient-boosted tree approaches for wave overtopping discharge prediction of coastal structures with composite slopes "without a berm". This study is expected to provide valuable, up-to-date information for estimating wave overtopping risk, forecasting wave overtopping for warning and emergency evacuation of people in the event of extreme waves, risk minimization, and economic assessment of coastal protection projects.

2 Methodology

In this study, the support vector machine and gradient-boosted tree approaches were used to predict wave-overtopping discharge at coastal structures featuring composite slopes "without a berm". A flowchart of the research steps is shown in Figure 1.

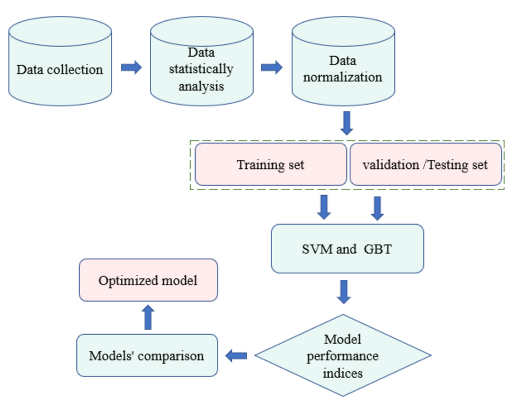

Fig 1. Flowchart of the SVM and GBT models used for wave- overtopping prediction

2.1 Data

The new, expanded database currently contains more than 17,942 tests, approximately 13,500 of which are solely for wave overtopping. About 10,000 schematized experiments on wave overtopping discharge q were collected from all over the world for the original CLASH database. The data used is the data that is convened for and assigned for the training of the soft computing methods, which were 4737 tests. Figure 2 illustrates the physical meaning of the parameters impacting coastal structures. The 13 parameters were analyzed through statistical measurements to summarize the characteristics of a data set, as shown in Table 1. Dikes, rubble mound breakwaters, berm breakwaters, caissons, and combinations of these structures are among the configurations tested, leading to intricate geometries.

Each test underwent a thorough screening process, during which a reliability factor, abbreviated RF, and a complexity factor, abbreviated CF, were assigned to each test (based on the volume and accuracy of the data). How simple it is to schematize the structure geometry using various geometrical factors determines the complexity factor. Reliable information or simple geometries received a score of 1, while less reliable information or complex geometries received a value of 3. A score of 4 indicates that the geometry was either too complex to be schematized or that the data were not trustworthy enough to be used.

The database was first expanded by adding already-existing databases on wave transmission and wave reflection. By keeping the same geometrical characteristics and pertinent climate parameters originally determined inside the CLASH project, the data assemblage was accomplished. Panizzo and Briganti[9] used the DELOS database as the basis for their testing, and Zanuttigh and Van der Meer used that same database for their experiments on the reflection coefficient (Kr)[10]. When wave overtopping or transmission values were provided, the reflection coefficient was also obtained. All the original data were once again checked for any potential missing values of the reflection coefficient.

2.2 Methods

The methods used in this study are Gradient Boosted Trees (GBT) and Support Vector Machine (SVM). The GBT is boosting method creates base models consecutively in contrast to bagging. By focusing on these tough to estimate training events, multiple models are developed sequentially to increase prediction accuracy. During the boosting process, examples that are challenging to estimate with the prior base models show up in the training data more frequently than examples that can be accurately estimated. Every new base model aims to fix the errors created by the previous base models. The boosting strategy was initially devised in response to Kearns' question (8) as follows (9) is one strong learner the same as a group of weak learners? A weak leaner is an algorithm that just marginally outperforms random guessing; a strong base model is a more accurate prediction or classification method that outperforms random guessing. Is unjustifiably connected to the problem? This question's answer is very important. A weak model is frequently easier to estimate than a strong one. Schapire provides evidence that the answer is yes by integrating multiple poor models into a single, very accurate model using boosting methods. The main distinction between boosting and bagging techniques is how the former carefully resamples the training data to deliver the most pertinent data for each succeeding model.

Each training step's modified distribution is dependent on the error that the previous models had caused. The likelihood of choosing a certain sample is not equal for the boosting algorithm, in contrast to the bagging approach, which uniformly selects each sample to create a training dataset. Misclassified or overestimated samples are more likely to be chosen with greater weight. As a result, each newly developed model emphasizes the samples that prior models incorrectly categorized. Boosting fits extra models that minimize a specific loss function, such as a squared error or an absolute error, averaged over the training data. The loss function calculates how much the true value differs from the projected value. The use of a forward stage-wise modeling approach is one of the approximations to this problem.

With the forward stage-wise technique, new base models are sequentially added without altering the parameters or coefficients of previously added models. The boosting approach is a type of "functional gradient decent" in terms of regression problems. By introducing a base model at each step that best reduces the loss function, it is an optimization strategy that minimizes a certain loss function. The Vapnik-developed SVM technique (1995). The SVM model was initially created to solve pattern recognition issues. SVM has recently been expanded to address non-linear regression estimation and time series prediction thanks to the advent of the -insensitive loss function. Based on the structural risk minimization concept, which is a strategy for minimizing the upper limit risk functional associated to generalization performance, SVM are efficient machine-learning techniques. An SVM is essentially a mathematical object, a technique (or recipe) for optimizing a specific mathematical function in relation to a specific set of data. However, the fundamental concepts behind the SVM algorithm may be described without ever reading an equation.

In fact, all that is required to comprehend the essence of SVM classification are four fundamental ideas: the soft margin, the maximum-margin hyperplane, the separating hyperplane, and the kernel function.

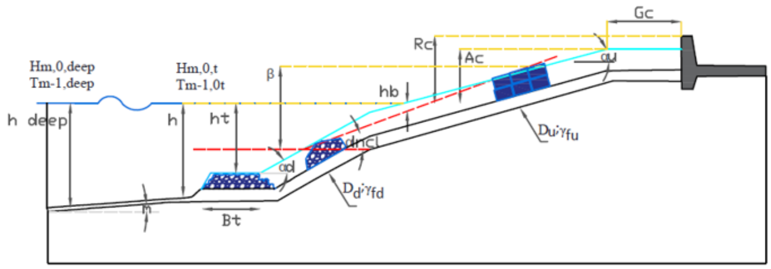

Fig 2. Cross section with application parameters for soft computing methods

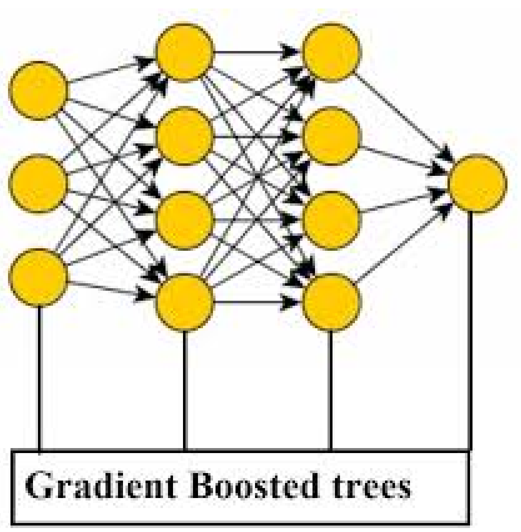

Fig 3. Explanation of the size and shape of GBT models

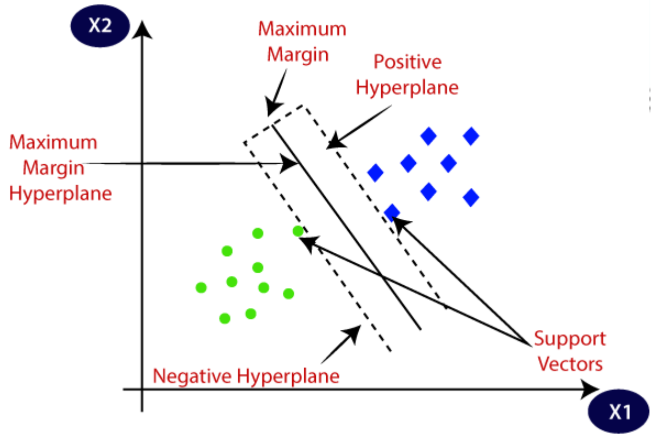

Fig 4. A decision boundary or hyperplane is used to classify two distinct categories in the figure

Table 1. Summary of statistical for the dimensional basic parameters of the used dataset

Parameter	N	Units	Minimum	Maximum	Mean	SD
$H_{m0\,toe}/L_{m1,0t}$	4737	[m]	0.017	1.480	0.127	0.077
β	4737	[°]	0.000	80.000	3.720	11.499
$h/L_{m1,0t}$	4737	[m]	0.029	5.010	0.462	0.399
$h_t/H_{m0\,toe}$	4737	[m]	0.029	5.010	0.440	0.403
$B_t/L_{m1,0t}$	4737	[m]	0.000	2.031	0.053	0.133
$R_c/H_{m0\,toe}$	4737	[m]	0.000	2.500	0.170	0.152
$A_c/H_{m0\,toe}$	4737	[m]	-0.030	2.500	0.161	0.153
$G_c/L_{m1,0t}$	4737	[m]	0.000	1.000	0.119	0.159
$\cot\alpha_d$	4737	[-]	0.000	7.000	2.307	1.193
$\cot\alpha_{incl}$	4737	[-]	0.000	7.000	2.340	1.203
$D/H_{m0\,toe}$	4737	[m]	0.000	0.109	0.024	0.026
γ_f	4737	[-]	0.380	1.000	0.722	0.276
Spread s	4737	[-]	0.000	10.000	0.346	1.400

3 Results and Discussion

Given that some of the data comes from small-scale models and other data comes from full-scale prototypes, the new EurOtop database advises against utilizing basic parameters as input to the models. Therefore, to avoid the wide range in raw parameter values, the basic data should be dimensionless. Dimensionless parameters are used to improve the accuracy and dependability of models.

The wavelength (Lm1,0t) can be calculated by using the following equation:

$$Lm1,0t = 1.56\,Tm1,0t^2 \tag{1}$$

The non-dimensional wave overtopping rate Sq is given by:

$$Sq = \frac{q}{\sqrt{g\,Hm0\,toe^{\,3}}} \tag{2}$$

$\cot\alpha_d$, $\cot\alpha_{incl}$, spread s, β, h / Lm1, 0t, Hm,0,t / $Lm1,0t$, h_t / Hm,0,t, B_t / $Lm1,0t$, γ_f, D / Hm,0,t, R_c / Hm,0,t, A_c / Hm,0,t, and G_c / $Lm1,0t$ are the final input and output dimensionless parameters used in creating soft computing models.

The mean square error (MSE), the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the correlation coefficient (R), the coefficient of performance (COP), the average absolute error (AAE), the scatter index (SI), the root mean square percentage error (RMSPE), and the Willmott index were used to evaluate the performance of the models (WI). The following are the equations for various statistical indicators:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Sq_{meas} - \left(Sq_{pred}\right|^{2}\right. \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Sq_{meas} - \left(Sq_{pred}\right|^{2}\right.} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(Sq_{meas} - \left(Sq_{pred}\right|\right. \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|Sq - Sq_p\right| \times 100 \tag{6}$$

$$R = \frac{\sum_{i=1}^{n}\left(Sq_{meas} - \overline{Sq_{meas}}\right)\left(Sq_{pred} - \overline{Sq_{pred}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Sq_{meas} - \overline{Sq_{meas}}\right)^2 \sum_{i=1}^{n}\left(Sq_{pred} - \overline{Sq_{pred}}\right)^2}} \tag{7}$$

$$COP = \frac{\sum_{i=1}^{n} Sq_{pred}}{\sum_{i=1}^{n} Sq_{meas}} \tag{8}$$

$$AAE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Sq_{meas} - (Sq_{pred}|}{Sq_{meas}}\right) \tag{9}$$

$$SI = \frac{RMSE}{\overline{Sq_{\mathrm{meas}}}} \tag{10}$$

$$RMSPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|Sq_{\mathrm{meas}} - Sq_{\mathrm{pred}}\right|}{Sq_{\mathrm{meas}}}\right)^2 \tag{11}$$

$$WI = 1 - \frac{\sum_{i=1}^{n}\left|Sq_{\mathrm{meas}} - Sq_{\mathrm{pred}}\right|^2}{\sum_{i=1}^{n}\left(\left|Sq_{\mathrm{meas}} - \overline{Sq_{\mathrm{meas}}}\right| + \left|Sq_{\mathrm{pred}} - \overline{Sq_{\mathrm{meas}}}\right|\right)^2} \tag{12}$$

Where Sq_{meas} and Sq_{pred} are the dimensionless measured and predicted values, n is the number of the observations, and $\overline{Sq}_{meas}$ and $\overline{Sq}_{pred}$ are respectively the average of Sq_{meas} and Sq_{pred}.

The SVM configuration parameters were utilized to create the training. Statistical indicators showed that the points were more concentrated and less evenly distributed, giving the impression that the SVM outputs were insufficient to predict the overtopping rates as depicted in Figure 5. The MSE, RMSE, MAE, and MAPE, in that order, were 0.000019, 0.0043, 0.0017, and 0.17. R stood at 0.94, while COP was at a respectable 1.00. AAE, SI, and RMSPE were at respective values of 11.69, 0.82, and 3870,11%. WI stood at 0.97. Figure 6 demonstrated that SVM could not predict overtopping rates more accurately than GBT.

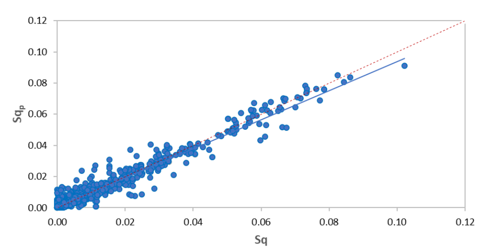

Fig 5. Comparison between the measured and predicted dimensionless overtopping discharge by SVM model

Figure 7 shows the results of the GBT model; the forecast points are scattered about the optimum line with some dots that were completely overstated. MSE, RMSE, MAE, and MAPE were determined by the statistical indices to be 0.000004, 0.002, 0.0011, and 0.11, respectively. COF was 1.00, while R was regarded as being high at 0.99. At 3.62, 0.39, and 1990,16%, respectively, AAE, SI, and RMSPE all performed admirably. At 0.99, WI was discovered. Figure 8 demonstrated that, with the smallest absolute error, the predicted and observed positions were nearly equivalent.

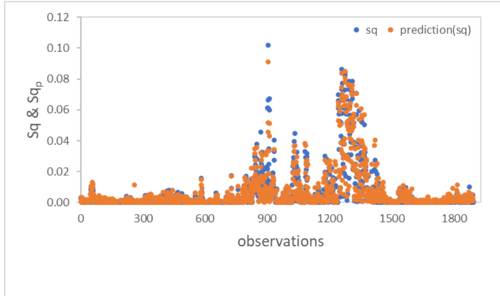

Fig 6. Error between the measured and predicted dimensionless overtopping discharge by SVM model

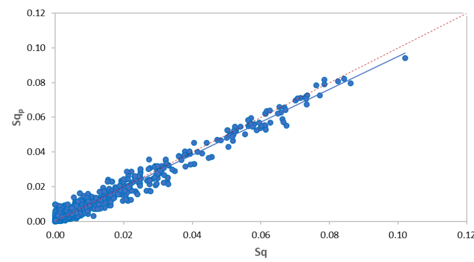

Fig 7. Comparison between the measured and predicted dimensionless overtopping discharge by GBT model

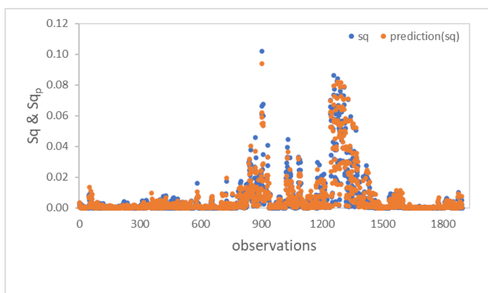

Fig 8. Error between the measured and predicted dimensionless overtopping discharge by GBT model

3.1 Comparison Between the SVM and GBT

In this section, the SVM and GBT algorithms' propensities for prediction were contrasted. The models' accuracy metrics are reported in Table 2. As can be observed from the table, the SVM models' matching MAPE was roughly 0.18, whereas the SVM's RMSE was 0.0049, resulting in a somewhat higher proportion of uncertainty compared to GBT. As can be seen in Figure 9, there is a slight correlation (R = 0.92) between the measured and predicted SVM values. Additionally, a WI (0.96) showed a poorer agreement between the measured and predicted values compared to GBT data.

When compared to an SVM model, the GBT model performed significantly better. Since RMSE and MAPE were 0.003 and 0.125, respectively, the model's error levels were reduced. The correlation coefficient would be nearly one.

As seen in Figure 9, the correlation coefficient is nearly one, which indicates improved performance. As a result, the AAE was lower at 5,06, and the SI was as well, at 0.57. This shows that the GBT method outperforms the other approaches analyzed in the study in terms of prediction estimates. The WI was close to one at 0.985 in Figure 9, demonstrating that this model was successful in fitting the data.

Since the SVM model was slightly off from the actual and predicted dash line in Figure 10, it generally tends to underestimate the overtopping rates, whereas the GBT model was closer to the dash line, indicating satisfactory performance of this model.

Figure 11 displays the discrepancy ratios for each method given the information. By plotting the histogram of DR values, DR is a technique for evaluating the qualitative proximity of forecasts to observations. The more closely the measured and predicted

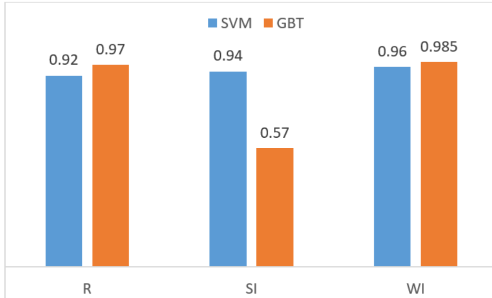

Fig 9. Comparison between R and WI values for the models

Table 2. Performance comparison between SVM and GBT for the coastal structure under study

Model	MSE	RMSE	MAE	MAPE	R	COP	AAE	SI	RMSPE %	WI
SVM	0.000024	0.0049	0.0018	0.18	0.92	0.996	11.8	0.94	3889.6	0.96
GBT	0.000009	0.003	0.00125	0.125	0.97	1.03	5.06	0.57	6313.4	0.985

overtopping rates match, the closer the DR values will be to zero.

Figure 11 shows clearly that 75% of the DR values for the GBT model were close to zero. However, the SVM only matches the measured and predicted values 66% of the while. This indicates that the GBT approach will provide more accurate and reliable estimates than the SVM method when used to forecast wave overtopping discharges.

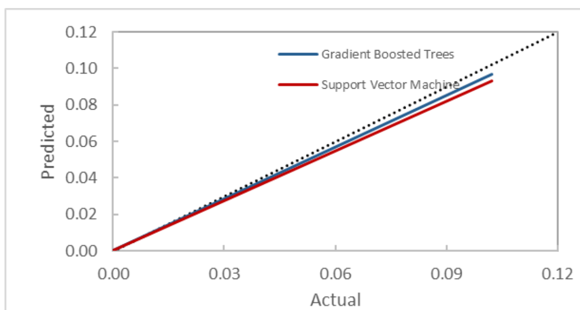

Fig 10. A scatter plot of the wave overtopping discharge levels for the SVM and GBT methods that were observed and predicted

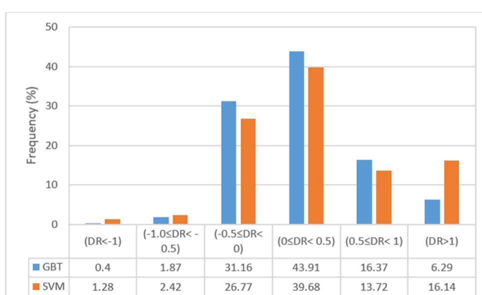

Fig 11. Comparison of discrepancy ratio values

To assess the validity of the GBT model, we compared its performance with that of the ANN model proposed by van Gent et al. (2007)[11]. The lower root mean squared error (RMSE) values attained by the GBT model during the analysis demonstrated that it outperformed the van Gent et al. model. Specifically, the RMSE of the GBT model was 0.003, whereas the RMSE of the van Gent et al. model was 0.29. These results indicate that the GBT model significantly reduced the overall error and accurately estimated the wave overtopping discharge. Overall, our findings suggest that the GBT model is a superior alternative to the van Gent et al. model for predicting wave overtopping discharge.

4 Conclusion

Coastal structures are primarily designed to prevent flooding and limit wave overtopping, but the ongoing effects of climate change, such as sea level rise and increased storm intensity and frequency, pose new challenges for the risk-based design of these structures. Accurately estimating overtopping discharges and understanding the characteristics of the overtopping flow over structures is crucial for ensuring the safety of people, activities, and goods in coastal areas or, at the very least, reducing their exposure to risk.

To address this challenge, the study utilized advanced machine learning techniques, specifically Support Vector Machine and Gradient Boosted Trees techniques, to predict wave overtopping discharge for a coastal structure with composite slopes ”without a berm”. The predictive performance of each model was evaluated using ten different parameters. The analysis of the EurOtop database (4737 data) found that the gradient-boosted trees technique produced exceptionally precise results in predicting wave overtopping discharge.

The analysis showed that the gradient-boosted trees model outperformed the ANN developed by van Gent et al. (2007) in terms of reducing prediction errors. This indicates that the GRNN model is more accurate and precise compared to other models.

However, further research is necessary to accurately represent more complex geometries, such as coastal structures with a berm, in the gradient-boosted trees. Additionally, we recommend comparing the performance of the gradient-boosted trees model with other available prediction methods for wave overtopping discharge to gain a better understanding of its effectiveness and applicability in different scenarios.

Acknowledgment

The authors would like to convey their sincere gratitude to the referees for their thorough evaluation and insightful criticism, both of which helped shape the manuscript into what it is today.

References

1) Pullen T, Allsop NWH, Bruce, Kortenhaus A, Schüttrumpf H, Van Der Meer JW. Manual on wave overtopping of sea defences and related structures . 2007. Available from: http://www.overtopping-manual.com/assets/downloads/EurOtop_II_2016_Pre-release_October_2016.pdf.
2) Verhaeghe H, De Rouck J, Van Der Meer J. Combined classifier–quantifier model: A 2-phases neural model for prediction of wave overtopping at coastal structures. *Coastal Engineering*. 2008;55(5):357–374. Available from: https://doi.org/10.1016/j.coastaleng.2007.12.002.
3) Molines J, Medina JR. Explicit Wave-Overtopping Formula for Mound Breakwaters with Crown Walls Using CLASH Neural Network–Derived Data. *Journal of Waterway, Port, Coastal, and Ocean Engineering*. 2016;142(3):142. Available from: https://doi.org/10.1061/(ASCE)WW.1943-5460.0000322.
4) Formentin SM, Zanuttigh BW, Van Der Meer JW. A Neural Network Tool for Predicting Wave Reflection, Overtopping and Transmission. *Coastal Engineering Journal*. 2017;59(1). Available from: https://doi.org/10.1142/S0578563417500061.
5) Bieman JPD, Van Gent MRA, Van Den Boogaard HFP. Wave overtopping predictions using an advanced machine learning technique. *Coastal Engineering*. 2021;166:103830. Available from: https://doi.org/10.1016/j.coastaleng.2020.103830.
6) Chen T, Guestrin C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016;p. 785–794. Available from: https://doi.org/10.1145/2939672.2939785.
7) Habib MA, O'sullivan J, Salauddin M. Comparison of machine learning algorithms in predicting wave overtopping discharges at vertical breakwaters. Copernicus GmbH. 2022. Available from: https://doi.org/10.5194/egusphere-egu22-329.
8) Elbisy MS. Machine Learning Techniques for Estimating Wave-Overtopping Discharges at Coastal Structures. . Available from: https://www.sciencedirect.com/science/article/pii/S0029801823003566.
9) Panizzo A, Briganti R. Analysis of wave transmission behind low-crested breakwaters using neural networks. *Coastal Engineering*. 2007;54(9):643–656. Available from: www.delos.unibo.it.
10) Zanuttigh B, Van Der Meer JW. Wave reflection from coastal structures in design conditions. *Coastal Engineering*. 2008;55(10):771–779. Available from: https://doi.org/10.1016/j.coastaleng.2008.02.009.
11) Van Gent MRA, Van Den Boogaard HFP, Pozueta BR, Medina JR. Neural network modelling of wave overtopping at coastal structures. *Coastal Engineering*. 2007;54(8):586–593. Available from: https://doi.org/10.1016/j.coastaleng.2006.12.001.