EFFECTIVENESS OF DECISION TREE AND KERNEL-BASED MACHINE LEARNING ALGORITHMS IN PREDICTING OVERTOPPING RATES FOR SLOPING STRUCTURES

M. A. Habib[1], J. J. O'Sullivan[2], S. Abolfathi[3], and M. Salauddin[2*]

The increasing frequency of storm surges driven by erratic climatic patterns linked to climate change has intensified overtopping events at coastal structures. Machine Learning (ML) has emerged as a valuable tool in predicting overtopping, complementing traditional empirical, physical, and numerical methods. The study investigates the application of two DT-based algorithms, namely Random Forest (RF) and Gradient Boosted Decision Trees (GBDT), and three kernel-based algorithms, including Artificial Neural Network (ANN), Support Vector Machines Regression (SVR) and Gaussian Process Regression (GPR) for predicting overtopping at sloping breakwaters. Required pre- and post-processing steps, such as feature selection and hyperparameter tuning, were integrated using a methodological framework and the prediction results were evaluated in terms of statistical metrics. The GPR algorithm was found to be the most accurate of the five algorithms tested, with r^2, *R*, RMSE and MAE values of 0.87, 0.94, 0.00031 and 0.23, respectively. Computational efficiency of the GBDT model was the highest among all the models. The physical consistency of the models was reported by applying feature importance analysis and the results were in agreement with the physical process of overtopping. The findings of this study can be beneficial for coastal engineers for fast and informed prediction tasks of wave overtopping at sloping structures.

Keywords: Wave Overtopping; Machine Learning; Decision Trees; Artificial Neural Networks; Random Forest; Gradient Boosted Decision Trees; Support Vector Machine Regression; Gaussian Process Regression

INTRODUCTION

Overtopping is the process of water movement by wave action from the seaward side of a coastal defense to the surrounding hinterland areas. Building resilience to wave-induced coastal flooding requires the accurate prediction of overtopping at coastal defense structures. In the context of an increased frequency of extreme storm events from climate change, it is imperative that methods for predicting wave overtopping become increasingly robust and utilize modern tools and techniques (Dong et al., 2020; 2024). Mean wave overtopping discharge per meter width of the structure, 'q', is a critical characteristic in assessing the severity of overtopping at sea defenses. Our understanding of the parameters that influence 'q' at coastal structures of varying geometry, including sloping breakwaters, are typically underpinned by any or a combination of numerical simulations, physical modelling and empirical methods.

The EurOtop (2018) overtopping manual serves as a comprehensive design and assessment guide for overtopping predictions of coastal defences, including seawalls, rubble mound breakwaters, and dikes, tailored for different geometries. The manual is principally based on nearly 18,000 tests collected from both field observations and laboratory studies on overtopping at coastal protections. The most recent edition of the manual, along with the accompanying dataset, was released in 2018 to assist coastal engineers worldwide in designing robust sea defences (see, EurOtop, 2018). For the prediction of mean overtopping rates at smooth sloping structures, EurOtop (2018) suggested a set of equations subjected to breaking and non-breaking waves (see Eqs. 1-2). Recent laboratory studies by Salauddin and Pearson (2020) on smooth sloping structures showed a strong agreement with the predictions from EurOtop (2018) empirical formulae.

For breaking waves ($\xi_{m-1,0} < 2$),

$$\frac{q}{\sqrt{gH_{m0}^3}} = \frac{0.023}{\sqrt{tan\alpha}}\,\xi_{m-1,0}\,\exp\left[-\left(2.7\frac{R_c}{\xi_{m-1,0}H_{m0}\gamma_f}\right)^{1.3}\right] \tag{1}$$

For non-breaking waves ($\xi_{m-1,0} > 2$),

[1]UCD Dooge Centre for Water Resources Research, School of Civil Engineering, University College Dublin, Dublin 4, Ireland
[2]UCD Dooge Centre for Water Resources Research, School of Civil Engineering, and UCD Earth Institute, University College Dublin, Dublin 4, Ireland.
[3]School of Engineering, University of Warwick, CV4 7AL, UK.
*Corresponding Author: md.salauddin@ucd.ie

$$\frac{q}{\sqrt{gH_{m0}^3}} = 0.09 \exp\left[-\left(1.5\frac{R_c}{H_{m0}\gamma_f}\right)^{1.3}\right] \tag{2}$$

where, q is the mean overtopping discharge, H_{m0} denotes the significant wave height, R_c is the crest-freeboard, γ_f is the influence factor for permeability and roughness of the slope, and $\xi_{m-1,0}$ is a wave breaker parameter calculated from Eq. 3:

$$\xi_{m-1,0} = \frac{\tan\alpha}{\sqrt{\frac{Hm_0}{L_{m-1,0}}}} \tag{3}$$

where, $L_{m-1,0}$ is the wavelength in deep water based on the spectral time period, $T_{m-1,0}(=\frac{gT_{m-1,0}^2}{2\pi})$.

With recent advances in computational efficiency, data-driven artificial intelligence (AI)) methods, such as machine learning (ML) algorithms, have shown potential for identifying complex underlying patterns in large datasets and have found application in methods to predict wave overtopping (Liu et al., 2020; Habib et al., 2022; Alvarellos et al., 2024). ML-based methods, including Artificial Neural Networks (ANNs), have also been endorsed for wave overtopping prediction in the EurOtop overtopping manual (Van Gent et al., 2007; Verhaeghe et al., 2008; Zanuttigh et al., 2016; Formentin et al., 2017). Additionally, recent studies reported the application of Decision Trees (DT) and Support Vector Machines Regression (SVMR) in predicting overtopping and scouring, showcasing their effectiveness and reasonable accuracy (den Bieman et al., 2021; Hosseinzadeh et al., 2021; Elbisy, 2023; Alshahri and Elbisy, 2023; Salauddin et al., 2023; Habib et al., 2024). While evidence of the successful application of ML algorithms for estimating wave overtopping rates at sea defences is increasing in the scientific literature (see, for example, Habib et al., 2023a, b; Elbisy 2023), the continuous development of increasingly robust AI tools means that there continues to be scope for improving the performance of ML algorithms for wave overtopping prediction tasks.

To this end, this study utilizes a systematic framework for the application of five ML algorithms in an overtopping study of simple sloped breakwaters. Three kernel-based methods SVMR, ANNs and GPR together with two DT methods (Random Forest (RF) and Gradient Boosted Decision Trees (GBDT)) were applied. In applying the models, the study focused on improving the accuracy of the methods through the reduction of redundant features by feature selection and by hyperparameter tuning of the model parameters. Interpretation of the key insights and results was achieved through a range of statistical metrics, and by feature importance analysis that considered the physical processes of wave overtopping as reflected in the ML methods.

MATERIALS AND METHODS

EurOtop Database

The data used in this study was derived from the EurOtop 2018 manual and includes results from physical model tests of overtopping for a wide range of geometries of coastal defense structures. The parameters and range of the dataset used to filter data related to sloping breakwaters are detailed in Table 1. The filtered data contained 1079 entries of overtopping data. Afterwards, feature selection (Sequential Forward Selection) was applied to the filtered data to extract the most impactful features in the dataset.

Table 1. Range of overtopping parameters used to extract data for simple sloped breakwaters

Feature	Description	Range of Values in Dataset
H_{m0d}	Significant wave-height in deep water [m]	0.02 to 0.3
T_{pd}	Time period of incident waves in deep water [s]	0.85 to 3.56
h	Water depth in front of the structure [m]	0.11 to 0.75
$H_{m0,toe}$	Significant wave-height at toe of the structure [m]	0.02 to 0.32
$T_{m,toe}$	Time Period at toe of the structure [s]	0.77 to 3.56
h_t	Water depth in front of the structure [m]	0.07 to 0.75

$\cot\alpha_d$	Cotangent of angle between structure slope downward berm and horizontal [-]	1.33 to 2
$\cot\alpha_u$	Cotangent of angle between structure slope upward berm and horizontal [-]	1.33 to 2
$D_{50,d}$	Nominal diameter of rock, downward slope [m]	0.026 to 0.1
$D_{50,u}$	Nominal diameter of rock, upward slope [m]	0.026 to 0.1
R_c	Crest freeboard [m]	0.001 to 0.37
B	Berm width [m]	0
γ_f	Permeability and Roughness Factor [-]	0.38 to 0.59; 0.55
$\tan\alpha_B$	Tangent of angle that sloping berm makes with horizontal [-]	0
G_c	Width of promenade [m]	0 to 0.875
RF	Reliability Factor [-]	1 to 3
CF	Complexity Factor [-]	1 to 3
FD	Freeboard Deficit [-]	0.5 to 1
$R_c/H_{m0,toe}$	Relative Freeboard [-]	0 to 2.83

Background of Machine Learning Algorithms

Random Forest and Gradient Boosted Decision Trees are two DT-based efficient machine learning techniques that are widely popular for regression tasks and are composed of ensembles of DTs (Neuwirth et al., 2020; Wang et al., 2020). RF and GBDT algorithms differ in the type of in-built features to reduce overfitting of the input data. The GBDT model implements gradient boosting, whereas the RF approach implements a bagging technique to reduce overfitting. The output quantity of an RF algorithm is computed from the average of all the predictions achieved by the individual DTs in the ensemble. The predictor model of an RF algorithm can be summarized by Eq. 4 (Rodriguez-Galiano et al., 2015).

$$\hat{f}_{rf(x)}^{K} = \frac{1}{K}\sum_{k=1}^{K} T(x) \tag{4}$$

where, $\hat{f}_{rf(x)}^{K}$ is the average predicted quantity from the RF function $rf(x)$ with x input vectors constituted from the features in a data set and K is number of Decision Trees, $T(x)$ in the ensemble.

Gradient Boosted Decision Trees is another algorithm based on ensemble DTs and designed for tackling classification and regression tasks in the field of data science GBDT incorporates gradient boosting, allowing nonlinear data classification, and regression tasks. Gradient boosting is a technique that computes the Mean Square Error (MSE) among the predicted and actual quantities which is then converted to a loss function. The loss function is then minimized by mathematical differentiation known as gradient descending in this case. GBDT, like RF, is popular for pattern recognition in high dimensional datasets exhibiting complex non-linear relationships (Wei, 2024).

SVMR is a derivative of support vector machines that can perform regression tasks by projecting input data to a set of support vectors. SVMR employs the Structure Risk Minimization Principle (SRM), that tends to reduce the maximum range of generalization errors by considering the total number of training errors and the confidence intervals. This approach is superior to the traditional empirical risk minimization principle that minimizes training error only (Ahmad et al., 2018). SVMR essentially transforms input data into higher-dimensional feature spaces using different kernel functions, that can tackle non-linearity in the data and perform regressions in the feature spaces (Li et al. 2009). The function for SVMR can be written as in Eq. 5.

$$f(x) = \sum_{i}^{l} y_i(\partial_i - \partial_i^{*})K(x_i, x) \tag{5}$$

where, $K(x_i, x)$ is the kernel function, l is the number of input data features and ∂_i, ∂_i^{*} are Lagrangian multipliers. In this study, the Gaussian Radial Basis Function (RBF) was adopted as the kernel function $K(x_i, x)$ as expressed in Eq.6.

$$K(x_i, x) = \exp(-\frac{\|x_i, x\|^2}{2\sigma^2}) \tag{6}$$

where σ is a kernel parameter.

ANNs are powerful tools for predicting wave overtopping at coastal structures. Researchers have relied on the application of ANNs to estimate wave overtopping discharge quantities with significant accuracy using experimental datasets such as the CLASH database (Formentin et al., 2017). Additionally, ANNs have been adopted to estimate wave transmission and reflection coefficients. The ability of ANNs to learn from large data, adaptability to particular datasets and producing outputs in a fast and efficient manner renders it as an indispensable tool in the application of artificial intelligence in coastal engineering (Raikar et al., 2018; Zanuttigh et al., 2016; Habib et al., 2023a). A feed forward and back propagation variant of ANN was implemented in this study, as shown in Fig. 1.

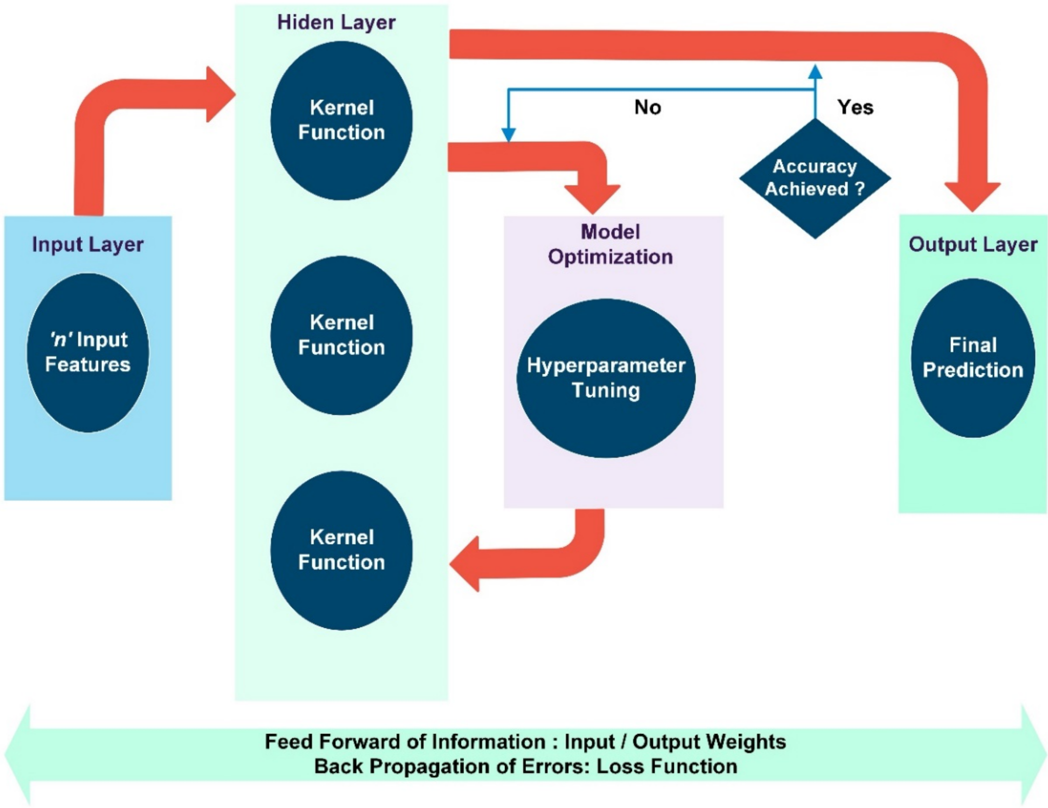

Figure 1. The schematic of a feed-forward and back propagation ANN algorithm adopted in this study (Adapted from Habib et al., 2023a).

The GPR is a regression-based non-parametric ML algorithm that is highly suitable for handling datasets with high dimensionality and non-linearity. The GPR algorithm is adaptive in nature from the context that instead of assuming a mapping function, it utilizes kernel functions to compute maximum and minimum values from the dataset by maintaining precision and flexibility (Komori et al., 2022). GPR acts similar to the SVR algorithm by converting input features into input vector $x \in R^d$ in a d-dimensional space. The output vectors are then mapped from the input vector utilizing a Gaussian Process. Application of GPR for overtopping prediction can be found in the work of Kim and Lee (2023) where, a co-variance function was developed to compute the relationship between the actual mean overtopping discharge and the residual differences between the actual and predicted quantities.

The optimum hyperparameters of all the algorithms were calculated using libraries from Scikit Learn. Following the pre-processing steps of data curation and feature selection, and the hyperparameter tuning of algorithms, the predicted overtopping quantities from the five algorithms were compared to the observed values in the EurOtop dataset. The effectiveness of the machine learning algorithms in predicting mean overtopping rates was evaluated by comparing the predictions to measured values using several statistical metrics. These metrics included the Coefficient of Determination (r^2), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). In Fig. 2, the workflow for data preparation, model development, and testing is summarized.

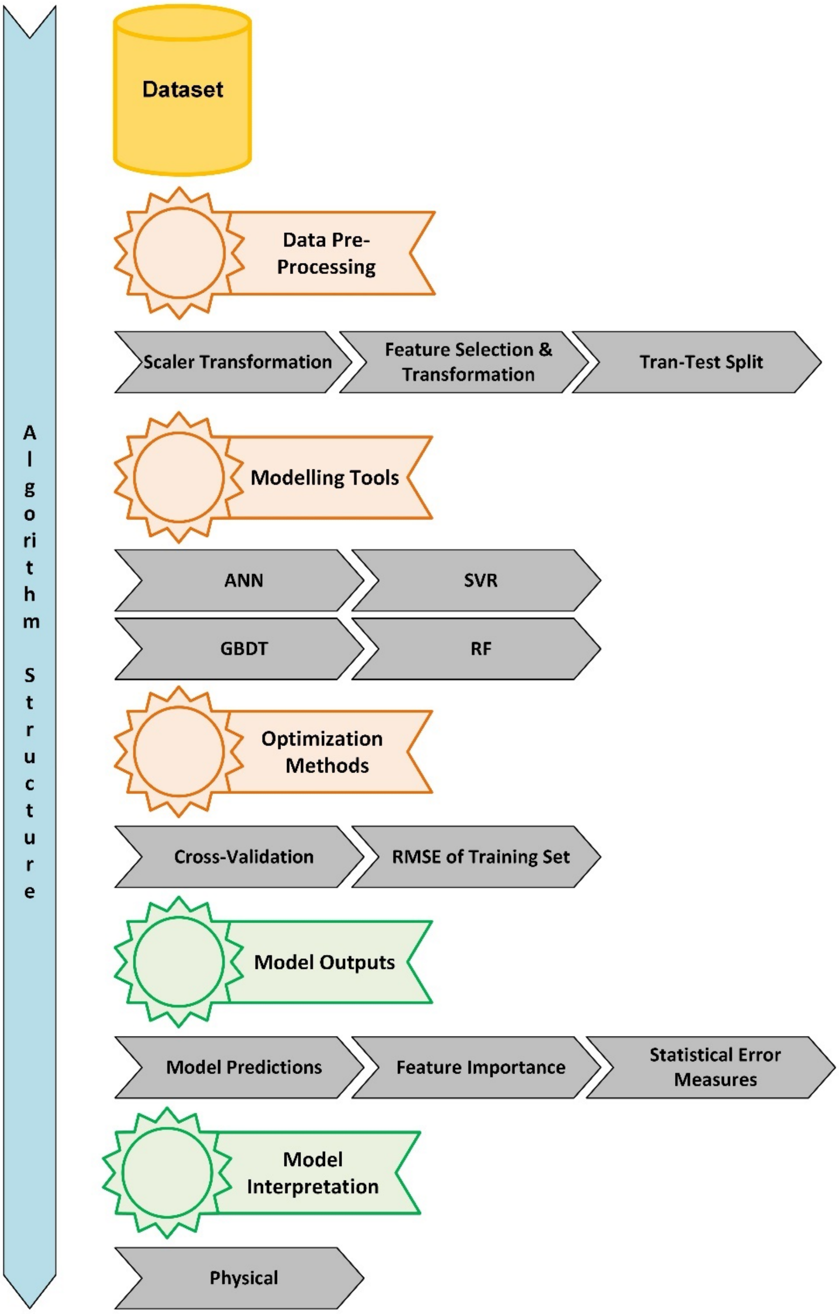

Figure 2. Flow chart of the adopted methodology in this study

RESULTS AND DISCUSSION

The comparison of the actual and predicted overtopping quantities (see Fig. 3) was investigated in terms of a Dimensionless quantity, referred to as Q ($=\frac{q}{\sqrt{9.81*(H_{mtoe})^3}}$), where '$q$' represents actual or predicted overtopping discharges. The approximate linearity of the graphs and the distribution of the data points indicates good agreement between values of Q based on actual and predicted overtopping quantities. The data in Fig. 3 also confirms that the tested ML algorithms provide comparatively better predictions for smaller, compared to larger, overtopping quantities.

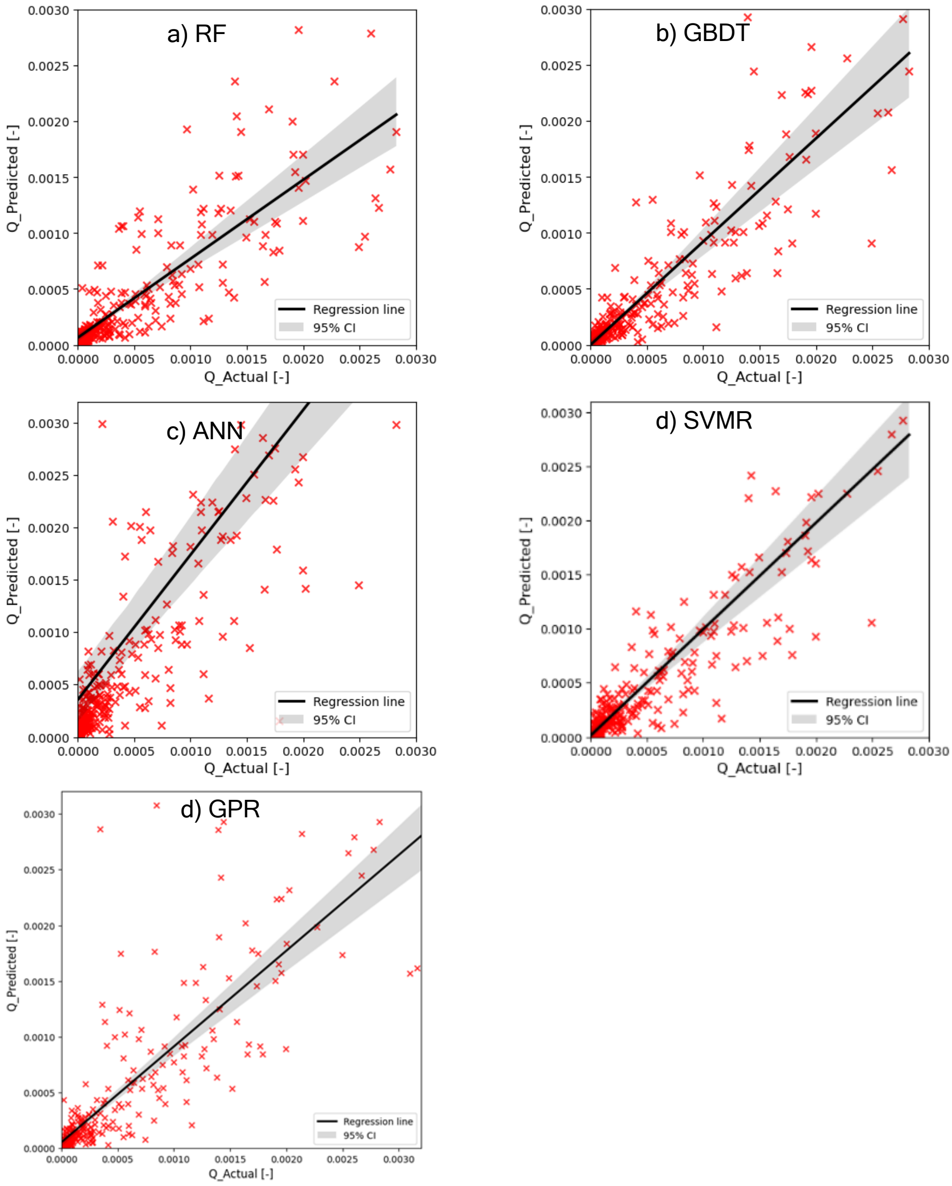

Figure 3. Comparison of predicted 'Q_predicted' ($=\frac{q\_predicted}{\sqrt{9.81*(H_{mtoe})^3}}$) and actual 'Q_actual' ($=\frac{q\_actual}{\sqrt{9.81*(H_{mtoe})^3}}$) overtopping quantities.

Furthermore, values of overtopping decreased with increasing crest-freeboard (crest freeboard = $R_c/H_{m,toe}$) and was consistent for all algorithms. The analysis of feature importance in all five tested algorithms confirmed crest-freeboard, R_c, to be the most impactful feature that influenced the overtopping discharge, '*q*', indicating that the tested ML models were capable of capturing the complexity of the physical processes that govern wave overtopping.

Based on the statistical error analysis of predicted and actual values, the GPR algorithm was found to be the most accurate of the five algorithms tested, with r^2, Pearson *R*, RMSE and MAE values of 0.87, 0.94, 0.00031 and 0.23, respectively (see Table 2). The GBDT model was shown to be the most computationally efficient algorithm, as evidenced by the lowest run time, in this instance being 85 s.

Table 2. Statistical metrics of the test-set in the ML models for predicting mean overtopping discharge					
Algorithm	**R^2**	**Pearson *R***	**MAE**	**RMSE**	**Computational Time (s)**
RF	0.80	0.90	0.30	0.00057	97
GBDT	0.85	0.93	0.28	0.00052	85
SVMR	0.81	0.91	0.30	0.00054	117
ANN	0.59	0.81	0.37	0.00071	128
GPR	0.87	0.94	0.23	0.00031	102

CONCLUSION

In this study, we assessed the application of two decision support tree-based models (RF and GBDT) and three kernel-based models (ANNs, SVMR and GPR) for predicting mean wave-overtopping rates at sloping structures. The machine learning models were trained using the EurOtop database with a 70%-30% train-test ratio. Hyperparameter tuning was conducted to optimize each algorithm for the specific dataset. Additionally, cross-validation was implemented to validate the training data before the algorithms were allowed to make predictions on the test set. Five statistical features were employed to evaluate the predictive performance of the tested machine learning models in estimating wave overtopping rates at sloped defences. Overall, it was observed that the tested ML methods provide consistent and reasonably accurate results in predicting overtopping volumes at sloping breakwaters. Among the machine learning models tested in this study, the GPR algorithm achieved highly accurate predictions of wave overtopping rates, making it the most effective algorithm for this task. Additionally, the decision tree (DT)-based algorithms outperformed the kernel-based methods in terms of computational efficiency. This can be attributed to the ability of DT-based algorithms to handle non-linear data more effectively than the kernel-based models.

ACKNOWLEDGMENTS

Md Arman Habib was awarded a grant by the Marine Institute, Ireland (NET-24-096) to present this work at the 38th International Conference on Coastal Engineering (ICCE) 2024. With thanks, the authors receive the EurOtop (2018) database from Professor Jentsje W. Van der Meer, Professor Barbara Zanuttigh, and Dr Sara Mizar Formentin.

REFERENCES

Ahmad, M. W., Reynolds, J., & Rezgui, Y. 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees, *Journal of Cleaner Production*, 203, 810–821. https://doi.org/10.1016/j.jclepro.2018.08.207

Alshahri, A.H. and Elbisy, M.S. 2023. Assessment of Using Artificial Neural Network and Support Vector Machine Techniques for Predicting Wave-Overtopping Discharges at Coastal Structures, *Journal of Marine Science and Engineering*, 11(3), p.539.

Alvarellos, A., Figuero, A., Rodríguez-Yáñez, S., Sande, J., Peña, E., Rosa-Santos, P. and Rabuñal, J. 2024. Deep Learning-Based Wave Overtopping Prediction, *Applied Sciences*, 14(6), p.2611.

den Bieman, J.P., van Gent, M.R. and van den Boogaard, H.F. 2021. Wave overtopping predictions using an advanced machine learning technique, *Coastal Engineering*, 166, p.103830.

Dong, S., Abolfathi, S., Salauddin, M. and Pearson, J.M., 2020. Spatial distribution of wave-by-wave overtopping at vertical seawalls, *Coastal Engineering Proceedings*, (36v). https://doi.org/10.9753/icce.v36v.structures.17

Dong, S., Salauddin, M., Abolfathi, S. and Pearson, J.M., 2024. Improved prediction of wave overtopping rates at vertical seawalls with recurve retrofitting, *Ocean Engineering*, *302*, p.117647. https://doi.org/10.1016/j.oceaneng.2024.117647

Elbisy, M. 2023. Machine learning techniques for estimating wave-overtopping discharges at coastal structures, *Ocean Engineering*, *273*, 113972. https://doi.org/10.1016/j.oceaneng.2023.113972

EurOtop 2018. Manual on wave overtopping of sea defences and related structures. An overtopping manual largely based on European research, but for worldwide application. Van der Meer, Allsop, Bruce, De Rouck, Kortenhaus, Pullen, Schüttrumpf, Troch and Zanuttigh. www.overtopping-manual.com

Formentin, S. M., Zanuttigh, B., & Van der Meer, J. W. 2017. A Neural Network Tool for Predicting Wave Reflection, Overtopping and Transmission, *Coastal Engineering Journal*, *59*(1). https://doi.org/10.1142/S0578563417500061

Habib, M.A., O'Sullivan, J.J. and Salauddin, M. 2022. Prediction of wave overtopping characteristics at coastal flood defences using machine learning algorithms: a systematic rreview. Proceedings of the 9th International Conference on Coastal and Ocean Engineering (ICCOE 2022), IOP Conference Series: Earth and Environmental Science (Vol. 1072, No. 1, p. 012003), IOP Publishing. DOI 10.1088/1755-1315/1072/1/012003

Habib, M. A., O'Sullivan, J. J., Abolfathi, S., & Salauddin, M. 2023a. Enhanced wave overtopping simulation at vertical breakwaters using machine learning algorithms, *PLOS ONE*, 18(8), e0289318. https://doi.org/10.1371/journal.pone.0289318

Habib, M.A., Abolfathi, S., O'Sullivan, J.J. and Salauddin, M. 2023b. Prediction of wave overtopping rates at sloping structures using artificial intelligence, *Proceedings of the 40th IAHR World Congress. Rivers–Connecting Mountains and Coasts*, IAHR, pp. 404-413. https://doi.org/10.3850/978-90-833476-1-5_iahr40wc-p0115-cd

Habib, M.A., Abolfathi, S., O'Sullivan, J.J. and Salauddin, M. 2024. Efficient data-driven machine learning models for scour depth predictions at sloping sea defences, *Frontiers in Built Environment*, 10, p.1343398. DOI: 10.3389/fbuil.2024.1343398

Hosseinzadeh, S., Etemad-Shahidi, A. and Koosheh, A. 2021. Prediction of mean wave overtopping at simple sloped breakwaters using kernel-based methods *Journal of Hydroinformatics*, 23(5), pp.1030-1049.

Komori, M., Teraji, T., Shiroshita, K., & Nittono, H. (2022). Examination of morphological traits of children's faces related to perceptions of cuteness using Gaussian process ordinal regression. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.979341

Kim, T., & Lee, W.-D. (2023). Prediction of wave overtopping discharges at coastal structures using interpretable machine learning. *Coastal Engineering Journal*, *65*(3), 433–449. https://doi.org/10.1080/21664250.2023.2233312

Li, Q., Meng, Q., Cai, J., Yoshino, H., & Mochida, A. 2009. Applying support vector machine to predict hourly cooling load in the building *Applied Energy*, *86*(10), 2249–2256. https://doi.org/10.1016/j.apenergy.2008.11.035

Liu, Y., Li, S., Zhao, X., Hu, C., Fan, Z. and Chen, S. 2020. Artificial neural network prediction of overtopping rate for impermeable vertical seawalls on coral reefs, *Journal of Waterway, Port, Coastal, and Ocean Engineering*, 146(4), p.04020015.

Neuwirth, C., Snyder, C., Kremser, W., Brunauer, R., Holzer, H., & Stöggl, T. 2020. Classification of Alpine Skiing Styles Using GNSS and Inertial Measurement Units, *Sensors*, *20*(15), 4232. https://doi.org/10.3390/s20154232

Raikar, R. V., Wang, C.-Y., Shih, H.-P., & Hong, J.-H. 2016. Prediction of contraction scour using ANN and GA, *Flow Measurement and Instrumentation*, *50*, 26–34. https://doi.org/10.1016/j.flowmeasinst.2016.06.006

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines, *Ore Geology Reviews*, *71*, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Salauddin, M. and Pearson, J.M., 2020. Laboratory investigation of overtopping at a sloping structure with permeable shingle foreshore, *Ocean Engineering*, 197, p.106866. https://doi.org/10.1016/j.oceaneng.2019.106866

Salauddin, M., Shaffrey, D. and Habib, M.A. 2023. Data-driven approaches in predicting scour depths at a vertical seawall on a permeable shingle foreshore, *Journal of Coastal Conservation*, 27(3), p.18. https://doi.org/10.1007/s11852-023-00948-w

Van Gent, M.R.A., Van den Boogaard, H.F.P., Pozueta, B., & Medina, J.R. 2007. Neural network modelling of wave overtopping at coastal structures, *Coastal Engineering*, *54*(8). https://doi.org/10.1016/j.coastaleng.2006.12.001

Verhaeghe, H., De Rouck, J., & Van der Meer, J. W. 2008. Combined classifier–quantifier model: A 2-phases neural model for prediction of wave overtopping at coastal structures, *Coastal Engineering*, *55*(5). https://doi.org/10.1016/j.coastaleng.2007.12.002

Wang, S., Zhuang, J., Zheng, J., Fan, H., Kong, J., and Zhan, J. 2021. Application of Bayesian Hyperparameter Optimized Random Forest and XGBoost Model for Landslide Susceptibility Mapping, *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.712240

Wei, S., and Lee, S. 2024. Financial Anti-Fraud Based on Dual-Channel Graph Attention Network, *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(1), 297–314. https://doi.org/10.3390/jtaer19010016

Zanuttigh, B., Formentin, S. M., and Van der Meer, J. W. 2016. Prediction of extreme and tolerable wave overtopping discharges through an advanced neural network, *Ocean Engineering*, *127*. https://doi.org/10.1016/j.oceaneng.2016.09.032