DEVELOPMENT OF XGBOOST MODEL FOR WAVE OVERTOPPING USING ENHANCED CLASH DATABASE

MOHAMED TAREK MOHAMED FOUAD

FACULTY OF ENGINEERING
UNIVERSITI MALAYA
KUALA LUMPUR

2024

DEVELOPMENT OF XGBOOST MODEL FOR WAVE OVERTOPPING USING ENHANCED CLASH DATABASE

MOHAMED TAREK MOHAMED FOUAD

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING SCIENCE

FACULTY OF ENGINEERING
UNIVERSITI MALAYA
KUALA LUMPUR

2024

UNIVERSITI MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Mohamed Tarek Mohamed Fouad

atric No: S2122509/1

Name of Degree: Master of Engineering science

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"): **Development of xgboost model for wave overtopping using enhanced clash database.**

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the Universiti Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 29-1-2024

Subscribed and solemnly declared before,

Witness's Signature Date: 30 January 2024

Name:

Designation

DEVELOPMENT OF XGBOOST MODEL FOR WAVE OVERTOPPING USING ENHANCED CLASH DATABASE

ABSTRACT

The accurate prediction of wave overtopping is crucial for designing resilient coastal structures. This thesis presents a comprehensive study on estimation of wave overtopping using (XGB) algorithm, with a focus on both model development and experimental validation. In the first part of the thesis, the focus was on the development of the XGB model for wave overtopping prediction. The methodology started with exploring the database parameters, followed by rigorous data preprocessing to ensure data quality. The model tuning process was elaborated, incorporating the utilization of hyperparameters to enhance predictive performance. After the preprocessing phase, the number of parameters chosen for the model development was 36 parameters, while the number of data points taken from the dataset was 5670 tests. The preprocessed database was split into 70% for training and 30% for testing in the XGB model. The model attained high predictive accuracy with RMSE of 0.28 m^3/s/m, a percentage error of 4.9%, and a high correlation coefficient (R) of 0.95. Percentage error was used as the primary error metric, underpinning its effectiveness in quantifying differences in prediction. The thesis examined model performance in different conditions by categorizing wave overtopping rate (q) data into low, medium, and high ranges. The low range consisted of 893 points while the medium and high range contained 772 and 36 points respectively. RMSE values for low, medium, and high q ranges were 0.34 m^3/s/m, 0.23 m^3/s/m, and 0.17 m^3/s/m, respectively. The percentage error statistics for these ranges were 4.9%, 4.9%, and 7.4%, respectively. Model validation is executed via the bootstrap resampling technique to reveal the model inherent robustness. Following the implementation of the resampling technique, the model showed a poorer result, with an RMSE of 0.31 m^3/s/m, an R value of 0.94, and a percentage error of 5.4%. To validate the performance of the model, the results were compared to an existing XGB model developed by Den Bieman (DB) that used the same database. Achieving similar results confirmed the good performance of the model and the XGB technique reliability. The second part of the thesis delved into the experimental aspect, contributing novel data to the existing database. A thorough designed experiment was conducted within the National Hydraulic Research Institute of Malaysia (NAHRIM), featuring comprehensive information about the wave flume, wave generator system, and data acquisition setup. The experimental design, encompassing wave conditions and data collection procedures, was outlined. Adding 49 new tests to the existing database had a small impact on predictive performance, with a percentage error of 10.09% for the original dataset and 10.43% for the updated dataset. The combination of model development and physical experiment contributed to a better understanding of wave overtopping phenomena. The results underscored the potential of the XGB algorithm in accurate wave overtopping prediction, while also emphasized the challenges and considerations when integrating experimental data into existing predictive frameworks.

Keywords: Wave Overtopping; Coastal Structure; Artificial Intelligence; Gradient boosting decision trees (XGBoost); Laboratory experiments

PEMBANGUNAN MODEL XGBOOST UNTUK GELOMBANG OVERTOPPING MENGGUNAKAN PANGKALAN DATA CLASH YANG DITINGKATKAN

ABSTRAK

Ramalan tepat perlebihan ombak adalah amat penting dalam kejuruteraan pantai dan pengurusan pantai, ini kerana ia dapat membantu dalam reka bentuk struktur pantai yang berdaya tahan. Kajian ini membentangkan kajian komprehensif tentang meramalkan perlebihan ombak menggunakan algoritma XGBoost (XGB), merangkumi pembangunan model dan pengesahan eksperimen. Bahagian pertama kajian memfokuskan kepada pembangunan model XGB untuk ramalan perlebihan gelombang. Metodologi bermula dengan penerokaan parameter pangkalan data secara meluas, diikuti dengan prapemprosesan data yang rapi untuk memastikan kualiti data. Proses penalaan model telah dihuraikan, menggabungkan penggunaan hiperparameter untuk meningkatkan prestasi ramalan. Set data latihan model XGB menggunakan 70% pangkalan data praproses, manakala set data ujian mengandungi 30% pangkalan data praproses. Model ini mencapai ketepatan ramalan yang luar biasa. dengan RMSE sebanyak 0.28 m^3/s/m, peratusan ralat sebanyak 4.9%, dan pekali penentuan tinggi (R) sebanyak 0.95. Penggunaan ralat peratusan dihujahkan sebagai metrik ralat utama, menyokong keberkesanannya dalam mengukur jurang ramalan. Untuk mendalami prestasi model, kajian ini membahagikan data perlebihan gelombang kepada julat kecil, sederhana dan tinggi, dan juga menjelaskan prestasi model merentas pelbagai keadaan. Nilai RMSE yang didapati adalah 0.34 m^3/s/m, 0.23 m^3/s/m, dan 0.17 m^3/s/m bagi julat q rendah, sederhana dan tinggi. Peratusan statistik ralat untuk julat ini ialah 4.9%, 4.9% dan 7.4%, setiap satunya. Pengesahan model dilaksanakan menggunakan teknik pensampelan semula bootstrap, agar dapat mendedahkan keteguhan model. Yang penting, model ini, apabila dilatih tanpa pensampelan semula, menunjukkan prestasi ramalan yang lebih baik berbanding pendekatan berasaskan pensampelan semula, dengan RMSE optimum 0.31 m^3/s/m, nilai R 0.94, dan peratusan ralat sebanyak 5.4%. Untuk mengesahkan prestasi model, hasilnya dibandingkan dengan model XGB sedia ada yang menggunakan pangkalan data yang sama, dan sedikit peningkatan yang ditunjukkan oleh model baharu. Bahagian kedua kajian ini menyelidik aspek eksperimen, menyumbang data baharu kepada pangkalan data sedia ada. Satu eksperimen yang direka dengan teliti yang telah dijalankan dalam Institut Penyelidikan Hidraulik Kebangsaan Malaysia (NAHRIM), yang menampilkan maklumat komprehensif tentang flume ombak, sistem penjana ombak dan persediaan pemerolehan data. Reka bentuk eksperimen yang merangkumi keadaan ombak dan prosedur pengumpulan data, telah digariskan. Walaupun usaha dalam mengumpul data baharu, pembesaran pangkalan data sedia ada dengan bilangan terhad sebanyak 49 ujian eksperimen menghasilkan impak yang sederhana terhadap prestasi ramalan dengan peratusan nilai ralat sebanyak 10.09% untuk set data asal dan 10.43% untuk set data selepas menambah data baharu. Ujian Pendekatan dwi pembangunan model dan pengesahan eksperimen memberi penerangan tentang kerumitan dan variasi fenomena ombak pantai. Hasilnya menggariskan potensi algoritma XGB dalam ramalan perlebihan ombak yang tepat, sambil menekankan cabaran dan pertimbangan apabila menyepadukan data eksperimen ke dalam rangka kerja ramalan sedia ada. Kajian ini bukan sahaja memajukan bidang kejuruteraan pantai tetapi juga memberikan pandangan berharga untuk usaha masa depan dalam permodelan ramalan dalam bidang proses persekitaran yang kompleks.

Kata kunci: Perlebihan Ombak, Struktur Pantai, XGBoost, Kecerdasan Buatan, Eksperimen Makmal

ACKNOWLEDGEMENTS

In the name of Allah, the Most Compassionate, the Most Merciful.

First and foremost, I would like to express my deepest graduate to my supervisors Dr. Muhammad Shazril and Prof. Ahmed Elsahfie for their generous guidance, persistent support, and invaluable insights throughout my study. The meetings and interesting discussions I had with you will always hold a special place in my memory. This thesis would not be what it is without your continuous advice and motivation.

I extend my sincere thanks to UM staff and my colleagues from the civil engineering department, for their friendship and support during my master degree. Special appreciation to Ahmed Alnahhal who have become more than a colleague for the exciting times we have shared and for his invaluable help and advice. I am very grateful to my peer and friend, Omar Momen at Heinrich Heine University in Germany for sharing his knowledge, valuable guidance and helping in carrying out the programming part. Also, I am grateful for my friend Ahmed Ismail, who has supported me throughout this journey with unwavering help.

I gratefully acknowledge the support from the Ministry of Science, Technology & Innovation (MOSTI) through grant bearing project number MOSTI001-2021TED1 (TDF05211386). Also, I would like to extend my gratitude to Dr. Zanuttigh for her help by providing the latest version of the CLASH Database.

I am deeply indebted to my parents and sister for their endless love, support, and encouragement throughout my life. Their sacrifices made the impossible, possible.

In token of my gratitude towards my parents and sister, I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

A_c	:	Armour crest freeboard of a structure
ANN	:	Artificial Neural Network
ANFIS	:	Adaptive Neuro-Fuzzy Inference System
B	:	Width of the berm of a structure, measured horizontally
B_h	:	Width of the horizontally schematized berm of a structure
B_t	:	Width of the toe of a structure
CF	:	Complexity factor of a structure section
f	:	Frequency
G_c	:	Crest width of a structure
h	:	Water depth at the toe of a structure
H_s	:	Significant wave height
h_b	:	Water depth on the berm of a structure
h_{deep}	:	Water depth at deep water
h_t	:	Water depth on the toe of a structure
H	:	Wave height
H_{m0}	:	Estimate of significant wave height from spectral analysis
$H_{m0\ deep}$	:	Estimate of significant wave height from spectral analysis at deep water
$H_{m0\ toe}$	:	Estimate of significant wave height from spectral analysis at the toe of a structure
H_{rms}	:	Root mean square wave height
L	:	Wave length
m	:	Measure of the slope of a foreshore = 1 (unit vertically) : m (units horizontally)
N_{ow}	:	Number of overtopping waves

N_w	:	Number of incident waves
q	:	Mean overtopping discharge per meter structure width
R_c	:	Crest freeboard of a structure [m] = height of ‘wave return point’ of a structure in relation to SWL
RMSE	:	Root Mean Square Error
R	:	Pearson Correlation Coefficient
R^2	:	Coefficient of determination
RF	:	Reliability factor of overtopping test = 1, 2, 3 or 4
SVM	:	Support Vector Machine
SWL	:	Still water level
T	:	Wave period = 1/f
T_m	:	Mean wave period derived either from time domain analysis or from spectral analysis
T_p	:	Peak wave period derived from spectral analysis
T_s	:	Significant wave period
PE	:	Percentage Error
V	:	Volume of overtopping wave per unit crest width
XGB	:	Extreme Gradient Boosting
α	:	Slope angle
α_B	:	Angle that sloping berm makes with a horizontal
α_d	:	Angle that structure part below the berm makes with a horizontal
α_{excl}	:	Mean angle that structure makes with a horizontal, excluding the berm
α_{incl}	:	Mean angle that structure makes with a horizontal, including the berm
ß	:	Angle of wave attack with respect to the normal on a structure

CHAPTER 1: INTRODUCTION

1.1 Background study

Coastal structures are built to safeguard densely populated coastal regions from the destructive forces of waves, storm surges, flooding, and erosion. The height of the structure's crest is a crucial factor in determining its protective capabilities. With the changing climate, sea levels are rising, and more intense storms are occurring, highlighting the importance of designing effective protective structures. The volume of seawater that spills over the crest, known as "wave overtopping," is a critical consideration in this regard (Figure 1.1).

The design of coastal structures should aim for an "acceptable" level of wave overtopping. What is deemed acceptable depends on socio-economic factors. Constructing tall coastal structures that prevent any overtopping is usually undesirable due to their exorbitant cost. Furthermore, these structures obstruct the scenic view of the sea, which is a significant tourist attraction with economic implications.

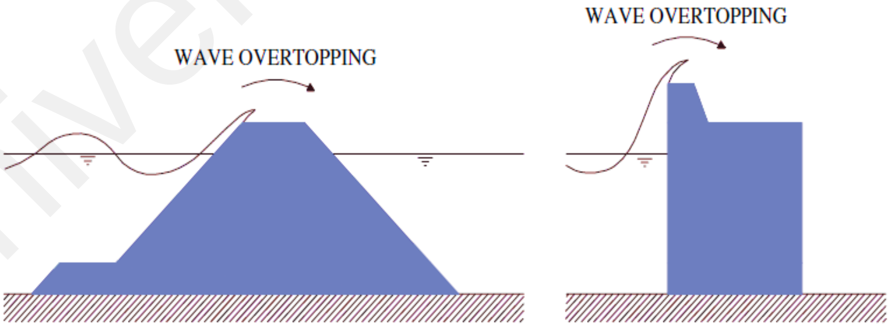


Figure 1.1: Wave overtopping: definition sketch

The study of overtopping phenomena commenced in the 1950s, with researchers like Thorndike Saville, (1986) pioneering the use of regular waves for overtopping tests. Since then, overtopping research has garnered significant attention, leading to the development of multiple models for predicting wave overtopping at various types of structures. The primary data for this study is derived from physical model trials, supplemented by prototype measurements. Initially, overtopping was solely replicated in laboratories using regular waves for several decades. However, later on irregular waves became the norm, which enhanced the accuracy of the prediction systems had been established. Notably, the well-known overtopping model based on irregular wave observations in the laboratory is the formula of (Owen, 1981), which remains influential to design structures until this day. In most published overtopping research, the focus lies on mean overtopping discharges, denoted as q, and expressed as flow rates per meter run of the defense structure (m3/s/m or l/s/m). Mean overtopping discharges are commonly used to set boundaries for allowable overtopping. Unlike the volume of individual overtopping waves, which may vary significantly, the mean overtopping discharge over approximately 1000 waves remain a stable parameter. However, it is essential to note that the local overtopping discharge from a single wave can be up to 100 times the observed time-averaged overtopping discharge during the storm peak due to the uneven distribution of overtopping in time and space (Verhaeghe, 2005). Early wave overtopping research focused primarily on specific structure types, resulting in overtopping models that were exclusively applicable to those particular structures. Vertical structures were often distinguished from sloping structural types (smooth or rough), and overtopping models were even developed for composite structure types. Empirical models used in prediction mean overtopping discharges have traditionally relied on a limited number of waves and structural factors. Consequently, each model is only valid for a specific type of structure. However, numerous studies have highlighted that several waves and structural variables influence overtopping. To address this limitation, many simple overtopping models now include adjustment variables. These correction factors account for extra overtopping influences that were not considered in the original models, such as oblique wave attack (Techniek, 2005). Engineers and coastal managers recognize that coastal defenses help mitigate the risk of wave overtopping, but it is essential to understand that seawalls do not always prevent overtopping completely; instead, seawalls reduce the occurrence. As depicted in Figure 1.2, waves can still overtop seawalls during storms, sometimes frequently and with considerable force.

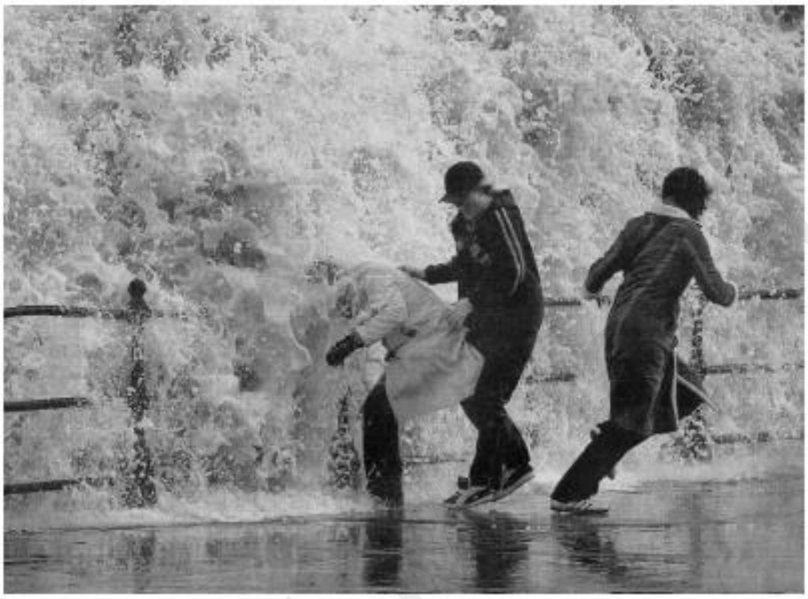


Figure 1.2: Waves overtop seawalls (Pullen et al., 2003)

The research project CLASH provided the foundation for the research that led to the conclusions and recommendations reported in this study. The CLASH project team included 13 members from seven different European countries. They came together primarily as a result of two observations. The initial observation was that there were few widely applicable and safe building design prediction methods. Also, it was showed that prediction approaches based on small-scale model data could be affected by scale or model effects. According to the study done by De Rouck et al., (2001), the 2 percent exceedance level for wave run-up on rubble mound slopes, recorded during full scale storms, was roughly 20% greater than modelled in small scale test facilities (scale 1/30). The CLASH project major goals were to deal with the problem of suspected scale effects for wave overtopping and to provide a generic wave overtopping prediction system that could be used for crest level design or assessment of coastal buildings (Geeraerts et al., 2007). Two primary steps have been taken to achieve these goals. The first step is based on wave overtopping prototype observations at selected field sites, laboratory replication, and numerical modelling. Three unique prototype measurements were taken in Europe: a steep rubble mound breakwater in Zeebrugge (Belgium), rubble mound breakwater in shallow water in Ostia (Italy) in addition to a vertical wall at Samphire Hoe (UK) (Pullen et al., 2003). The second phase involves compiling overtopping data from model test results from universities throughout the world into a single overtopping database. Initially, the CLASH database (STEENDAM et al., 2005) was utilized for neural network analysis. This database contained over 10,000 data entries derived from physical model experiments on different structures such as rubble mound breakwaters, berm breakwaters and dikes, which were collected from various institutions around Europe. The CLASH database consisted of 31 parameters, representing hydraulic, structural, and general parameters. Subsequently, Zanuttigh et al., (2016) extended the original database by including additional wave overtopping data. This expansion included more data on wave transmission and reflection, along with the addition of some parameters. The extended database (Eurotop, 2018), which builds upon the CLASH database, consists of over 17,000 tests with nearly 13,500 for wave overtopping only. This significant expansion brings the total number of parameters to 23 structural parameters, 13 hydraulic parameters, and 5 general parameters (Figure 1.3), resulting in an addition of 8 parameters compared to the original CLASH database (Zanuttigh et al., 2017). In this study, all parameters from the extended database were utilized and listed in Table 1.1. However, to obtain a suitable training dataset: 1) unreliable data was excluded, 2) weight factors were assigned to each data entry, and 3) scaling procedures were implemented. The Reliability Factor (RF) and Complexity Factor (CF) are two of the general parameters that are associated with test reliability and structural complexity. These parameters play a crucial role during both the dataset preprocessing and modeling phases. The RF ranges from 1 to 4, where RF value of 1 indicates a highly reliable test, while RF value of 4 suggests a test with low reliability. On the other hand, the CF is assigned a value of 1 for a simple structure, where the cross-section parameters precisely describe its characteristics. A CF value of 4, however, indicates a highly complex structure where an accurate description of the cross section is not possible. In terms of weighting each test differently during the modelling, van Gent et al., (2007) proposed a weighting factor (WF) formula, that will be incorporated in this study, and it is defined as

$$\mathbf{WF} = (\mathbf{4} - \mathbf{RF})(\mathbf{4} - \mathbf{CF}) \tag{1.1}$$

By using this formula, if RF = 1, which is a very reliable test, and CF =1, which indicates a simple structure, WF will have the values of 9. Hence, the most reliable and least complex data has the highest weight factor.

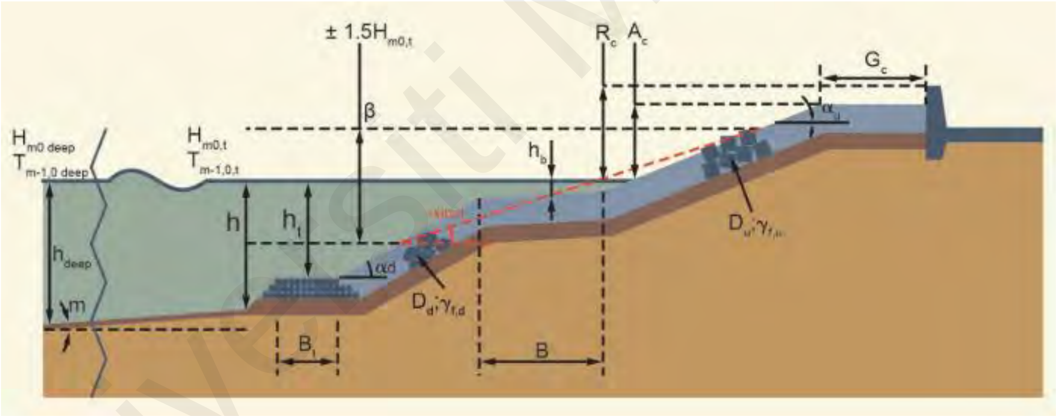


Figure 1.3: Simplification of the parameter definitions. (Eurotop, 2018)

Table 1.1: Summary of the features and its definition

#	Parameter	Unit	Definition of the parameter	Type
1	Name	-		general
2	$H_{m0\ deep}$	m	Off-shore significant wave-height	hydraulic
3	$T_{m\ deep}$	s	Off-shore average wave period	hydraulic
4	$T_{m-1,\ deep}$	s	Off-shore spectral wave period	hydraulic
5	$T_{p\ deep}$	s	Off-shore peak wave period	hydraulic
6	$h_{\ deep}$	m	Offshore water depth	structural
7	h	m	Water depth at the structure toe	structural
8	A_c	m	Wall height with respect to SWL	structural
9	β	°	Wave obliquity	hydraulic
10	m	-	Foreshore slope	structural
11	$H_{m0\ t}$	m	Significant wave-height at the structure toe	hydraulic
12	$T_{m\ t}$	s	Average wave period at the structure toe	hydraulic
13	$T_{m-1,\ t}$	s	Spectral wave period at the structure toe	hydraulic
14	$T_{p\ t}$	s	Peak wave period at the structure	hydraulic
15	$\cot\alpha_u$	-	Cotangent of the angle that the part of the structure below/above the berm makes with a horizontal	structural
16	$\cot\alpha_d$	-		structural
17	B_t	m	Toe width	structural
18	h_t	m	water depth on the toe of a structure	structural

Table1.1: Continued

19	$\cot\alpha_{incl}$	-	Cotangent of the mean angle that the structure makes with a horizontal, including/excluding the berm, in the run-up/run-down zone	structural
20	$\cot\alpha_{excl}$	-		structural
21	γ_{fd}	-	Roughness factor for $\cot\alpha_d$	structural
22	γ_{fu}	-	Roughness factor for $\cot\alpha_u$	structural
23	γ_f	-	Roughness factor [average in the run-up/down area in the new database	structural
24	Type	-	Type of structure and armor unit	structural
25	S	-	Spreading factor	structural
26	G_c	m	Crest width	structural
27	h_b	m	Berm submergence	structural
28	B_h	m	Horizontal berm width	structural
29	D	m	Average size of the structure elements in the run-up/down area	structural
30	D_u	-	Size of the structure elements along $\cot\alpha_u$	structural
31	D_d	-	Size of the structure elements along $\cot\alpha_d$	structural
32	B	m	Berm width	structural
33	P_{ow}	-	Percentage of the waves resulting in overtopping = (N_{ow} / N_w).100	hydraulic
34	$\tan\alpha_B$	-	Berm slope	structural
35	R_c	m	Crest height with respect to SWL	structural
36	K_t	-	(bulk) wave transmission coefficient	hydraulic

Table1.1: Continued

37	K_r	-	(bulk) wave reflection coefficient	hydraulic
38	CF	-	'Complexity Factor' The complexity of the test is indicated by a score with a possible range of 1 to 4.	general
39	RF	-	'Reliability Factor' The reliability of the test is indicated by a score with a possible range of 1 to 4.	general
40	q	$m^3/s/m$	Average specific wave overtopping discharge	hydraulic
41	Core data	-	Flag indicating the inclusion/exclusion from the core data of the ANN training	general

Machine learning encompasses a diverse range of algorithms that can automatically learn patterns and relationships from data, enabling predictive modeling and decision-making (Rogers, 2020). Some of the prominent machine learning techniques that have found applications in coastal engineering include supervised learning algorithms such as support vector machines (SVM), random forests, neural networks and recently extreme gradient boosting (XGB) (den Bieman et al., 2020). These techniques are widely used for coastal hazard prediction, storm surge forecasting, and shoreline change analysis, among others. Additionally, unsupervised learning algorithms like k-means clustering and hierarchical clustering have been instrumental in identifying coastal vulnerability hotspots and patterns in sediment transport.

The fundamental principle of extreme gradient boosting (XGB) lies in its utilization of the ensemble algorithm, which is built upon the gradient boosting tree (T. Chen & He, 2014). Gradient boosting, a prominent technique in ensemble algorithms (J. Friedman et al., 2000; J. H. Friedman, 2002), serves as the basis for XGB. XGB is an optimized implementation of the gradient boosting algorithm that has garnered widespread acclaim in industry and Kaggle machine learning competitions due to its exceptional efficiency. Similar to the gradient boosting decision tree (GBDT), XGB operates based on the principles of classification and regression tree theory (Ding et al., 2020; Le et al., 2019). To prevent overfitting, the optimized objective function of XGB introduces regularization terms (T. Chen & Guestrin, 2016), resulting in a composite objective function with two components. The first component measures the disparity between the predicted value and the actual value that representing the model deviation, while the second component represents the regularization term which responsible for capturing the variance of the control model (Zhou et al., 2021). The accuracy of the model predictions is influenced by both the deviation and variance of the model. Given a dataset D = $\{(x_i, y_i)\}$ consisting of n samples and m features, and a predictor composed of k base models, the predicted results for the samples can be expressed as

$$\hat{Y}_{\boldsymbol{i}} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{k}} \boldsymbol{f_k(x_i), f_k \epsilon \varphi} \tag{1.2}$$

$$\boldsymbol{\varphi = \{f(x) = w_s(x)\}(s: R^m \rightarrow T, w_s \epsilon R^T)} \tag{1.3}$$

where, 'xi' represents an individual sample, and for each sample, there is a prediction score denoted as 'fk(xi)'. The set φ represents a collection of regression trees, where each tree, denoted as 'f(x)', has its own structural parameters 's' and leaf weights 'w'. The variable 'T' represents the total number of leaves in a tree, while 'K' represents the number of trees used to combine or ensemble the results. Lastly, 'yi' refers to the predicted label associated with a particular sample. XGB algorithm works as decision trees, and the combination of these decision trees creates a more accurate and robust prediction model. The trees are grown sequentially, and each new tree is grown to correct the mistakes made by the previous trees as shown in Figure 1.4. This process continues until the desired level of accuracy is reached or until a specified stopping criterion is met which referred to as early stopping. Such features make XGB model particularly well-suited for large-scale and high-dimensional datasets, and can handle missing values, categorical variables, and unstructured data. Additionally, XGB has several built-in features that make it highly customizable, including support for parallel computing, a variety of evaluation metrics, and the ability to handle both regression and classification tasks. Note that this study utilizes the Python programming language and leverages the power of the XGBoost library in the implementation of the model.

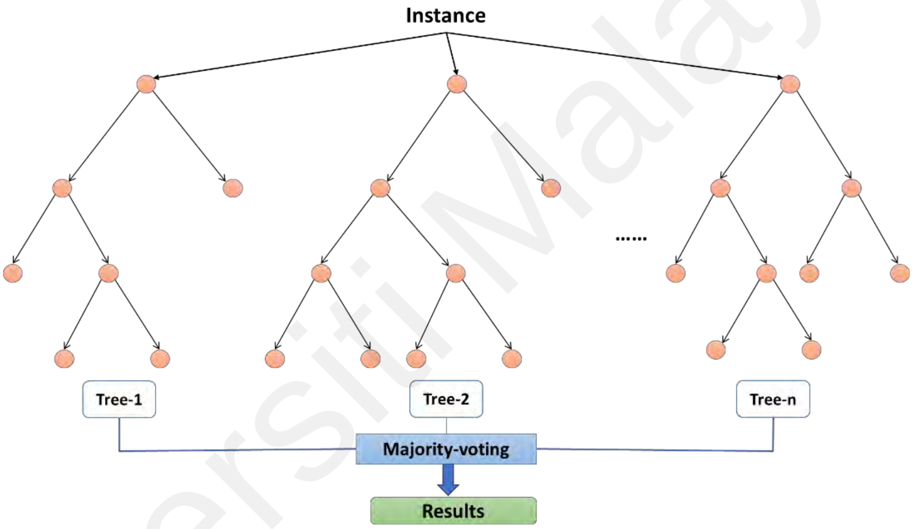


Figure 1.4: XGB Architecture

1.2 Problem Statement

Wave overtopping is a major issue in coastal engineering that occurs when waves exceed the height of a coastal structure, such as a sea wall, breakwater, or jetty, and spill over onto the land. This can cause significant damage to the structure and the surrounding area, i.e., coastal infrastructure, buildings, and roads. Overtopping can also pose a threat to human safety, particularly in densely populated coastal areas.

The prediction of wave overtopping is a complex problem that requires the integration of various factors, including wave height, wave period, wave direction, and the geometry and roughness of the coastal structure (K. K. Pillai et al., 2020). Therefore, reliable prediction of wave overtopping is crucial for the design of coastal structures, as it affects their stability and lifespan. The design of coastal structures must take into account the expected wave overtopping and ensure that the structure can withstand the forces generated by the overtopping waves (J. W. van der Meer et al., 2009).

Estimation of wave overtopping rate is typically carried out using mathematical models, such as empirical or semi-empirical models, or numerical models, such as wave tank or computational fluid dynamics (CFD) simulations. However, these models can be time-consuming and computationally intensive, and may not always accurately capture the complex/stochastic interactions between waves and coastal structures (Wee et al., 2021). In recent years, machine learning algorithms, such as Neural Network (NN), Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost), offer promising approaches for solving some complex engineering problems with high level of accuracy.

Application of machine learning in wave overtopping prediction has gained much attention in recent years and become an alternative to traditional mathematical models. Machine learning algorithms, such as XGBoost, can leverage large amounts of data to learn complex relationships between the various factors that influence wave overtopping. This allows for prediction of wave overtopping with high accuracy and speed, making it a valuable tool for coastal engineers (T. Chen & Guestrin, 2016).

Recent findings by den Bieman et al., (2020) and den Bieman et al., (2021), have provided strong evidence that XGBoost can serve as a robust alternative to traditional machine learning methods like NN and SVM. Their findings highlighted the growing recognition of XGBoost's potential to outperform and surpass older methodologies in various domains of machine learning. XGBoost, or simply XGB, is a powerful machine learning algorithm that has been widely used for various applications, including prediction and classification. XGB is an ensemble learning method that uses decision trees as base learners and incorporates various regularization techniques, such as shrinkage and random subsampling, to improve the accuracy of the model (Zhang et al., 2018). The XGB method is known to perform well on a wide range of problems and is suitable for large datasets, making it a promising approach for wave overtopping prediction (Lim & Chi, 2019).

1.3 Significant of research

The study holds significance for the field of coastal engineering by addressing the crucial need for accurate wave overtopping prediction, essential in designing resilient coastal structures. The study focuses on the development of a robust XGBoost (XGB) algorithm model, demonstrating high predictive accuracy with a low Root Mean Square Error (RMSE) of 0.28 m^3/s/m and a percentage error of 4.9%. The output variable (q) data was categorized into low, medium, and high ranges to enhance the comprehension of the database. This categorization not only facilitated a nuanced exploration of the dataset but also allowed for a discerning examination of how variations within these ranges influenced the model performance. The conduction of the physical experiment within the National Hydraulic Research Institute of Malaysia (NAHRIM) adds novel data to the existing database, contributing to a better understanding of wave overtopping phenomena. The combined insights from model development and experimental validation emphasize the potential of the XGB algorithm for accurate predictions and provide practical implications for coastal structure design.

1.4 Objectives of research

The main objectives of this research can be as follows: -

1. To review current research and models related to wave overtopping and dataset parameters.
2. To identify key hydrodynamic parameters in the EurOtop overtopping database.
3. To develop a reliable Extreme Gradient Boosting (XGBoost) model for estimation of overtopping rates using optimal hyperparameter values.
4. To assess the impact of additional new overtopping database on the performance of new Extreme Gradient Boosting (XGBoost) model.

1.5 Scope of work

This study is comprised of two distinct components. The first involves an in-depth examination of the Eurotop database, followed by the development of an AI model using XGB technique to predict wave overtopping. The second component involves conducting experimental work to measure wave overtopping, with the aim of expanding the existing database and assessing the impact of these new data on the original dataset.

1.6 Thesis layout

This thesis is structured to investigate the accurate prediction of wave overtopping in coastal engineering through the utilization of the XGBoost (XGB) algorithm. It contains five chapters, as follows:

Chapter 1 provides background information about wave overtopping and XGB algorithm. It provides the aims and objectives and the outline of the thesis.

Chapter 2, devoted to the literature review, presents an analysis of existing research on CLASH overtopping database with introducing the establishment of the dataset with the challenges and improvements provided. Then, the application of machine learning in wave overtopping is discussed with explaining the structure of each machine learning technique. Additionally, this chapter explains in detail the use of machine learning in coastal engineering with different machine learning techniques and it compares the results of each model. Lastly, it presents the advantages and disadvantages of each machine learning technique.

The methodology in chapter 3 outlines the research approach, divided into model development and experimental validation phases. It elaborates on steps such as database exploration, data preprocessing, hyperparameter tuning, model training, and experimental setup with data acquisition. In the model development section, it discusses the XGB model, including the optimization of hyperparameters, data partition, and application of standardization and scaling on input and output parameters. The subsequent section delves into the experimental validation, the experiment setup details in the National Hydraulic Research Institute of Malaysia (NAHRIM), wave conditions applied and empirical data collection.

Chapter 4 explores the practical testing and validation of wave overtopping prediction model using the XGBoost (XGB) algorithm. It examines the accuracy of the model in predicting wave overtopping. The model performance is assessed by comparing its predictions to actual observed data and using different metrics such as RMSE and PE to evaluate the performance and the accuracy of predictive models. Next, an investigation into how the model performs across different ranges of ‘q’ helps understand its adaptability to varying conditions and its reliability in different situations. The chapter proceeds to explore a technique called "Bootstrap Resampling" to validate the model consistency. This method assesses whether the model predictions remain reliable across different scenarios and whether they demonstrate robustness. Furthermore, a comparison between the newly developed model and an existing XGB model is conducted. This comparison aims to measure enhancements achieved in the model and offers insights into its predictive capabilities. Shifting focus to the physical experiment, the real-world conditions used in tests are examined. These conditions establish the basis for empirical data collection and provide context for the experiments. Then, an evaluation of the impact of new tests on the model predictions is undertaken. The incorporation of data from physical experiments helps gauge the model performance in real-world scenarios and assess whether the new data improves predictive accuracy of the existing dataset.

Finally, the key findings are highlighted, and recommendations for future studies are proposed.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The prediction of overtopping caused by wave run-up on beaches and structures is crucial for the design of coastal structures and the determination of flooding during normal storms and storm surge events (M. S. I. Ibrahim & Baldock, 2021). The height of the coastal structure is a very important element in the design phase, to obtain optimum design of coastal structures (Gallien et al., 2014). Also, the view of the sea cannot be blocked as it attracts tourists, and it should not be very low so it can protect the people, the vehicles, and the properties behind the structure (Chini & Stansby, 2012). Wave overtopping does not only lead to disasters like floods and erosion, but the event may also expose the people and the infrastructure to more severe hazards (Geeraerts et al., 2007). Therefore, the estimation of the overtopping volume and discharge is important for the coastal structures and to ensure the safety of the local stakeholders (Orme, 2015).

There are two common methods to measure the wave overtopping. The first method examines the volume per overtopping wave, whereas the second and most applied approach is the mean overtopping discharge over specific time intervals and per meter of structure width (Techniek, 2005). However, there is an absence of solid and powerful approach for prediction of wave overtopping at coastal structure. According to Geeraerts et al., (2007), a meaningful overtopping test requires at least 1000 waves. This is because wave overtopping is clearly undistributed in both time and space due to the irregular wave action. Therefore, it is not simple to measure the overtopping wave or to come up with a prediction method, thus designing the crest level of coastal structure. There are some formulas derived via empirical models, but it can only be implemented within a limited range, and it only covers a restricted number of structural designs (Verhaeghe, 2005). In addition, the precision of the overtopping measuring system is not specified, and it is occasionally impossible to determine whether a zero-q value indicates no overtopping at all or that the overtopping volume is too small. Although there is a huge number of tests and parameters in the CLASH database, only a small number of the data can be used for model training of a neural network. This is caused by the existence of the large number of tests with a low reliability factor (Zanuttigh et al., 2016a).

Soft computing refers to a collection of computational techniques that aim to mimic human-like reasoning and decision-making processes. These methods are designed to deal with complex and uncertain problems where conventional computing approaches may fall short (Dehghani et al., 2023). Soft computing methods typically include fuzzy logic, neural networks, genetic algorithms, and probabilistic reasoning. These methods are widely applied in various fields, including data analysis, pattern recognition, image processing, control systems, optimization, and decision support. Their flexibility, ability to handle uncertainty, and adaptability make them valuable tools for solving real-world problems where precise mathematical models may not be available or applicable (D. Ibrahim, 2016). With significant advances in computer technology over the past decade, artificial intelligence (AI) has been applied to environmental, maritime, and coastal issues, with many reliable and potential outcomes (S. H. Chen et al., 2008). Thus, it has become a possible and strong approach in model prediction. Artificial neural networks (ANNs), fuzzy logic, and hybrid systems, which are adaptive neuro-fuzzy inference system (ANFIS), are examples of AI techniques (Filippo et al., 2012). The artificial neural network (ANN) is a method of computing that simulates the biological neural network in the human brain. The ANN is composed of a sequence of nodes (neurons) organized in multiple layers. Each node in a layer receives and processes weighted input from the preceding layer before transmitting it to the output nodes through the next layer's links (Rezaie-Balf et al., 2019). Among ANN models, MLP-ANN is the most utilized network type. Figure 2.1 shows the MLP-ANN structure consisting of an input layer, one or more hidden layers and an output layer. Each layer is made up of an inter-connected collection of simple processing components called neurons. Layered structure is used to arrange these processing components. Each neuron in one layer is connected to the neutron in the next layer and weights are the connections between layers (Muslim et al., 2020).

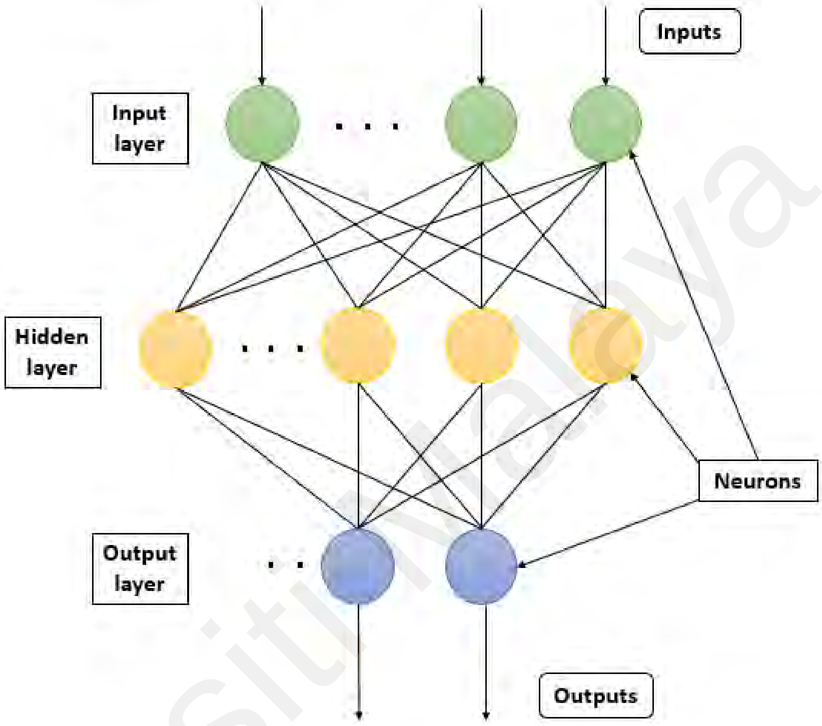


Figure 2.1: Multilayer perceptron neural network (MLP-ANN) structure

According to Figure 2.2, the application of machine learning on wave overtopping has gained attention in recent years, mainly to address the limited range of validity of empirical model related to CLASH database. Therefore, this Chapter reviews the challenges and problems mentioned above and considers steps to address them.

Table 2.1 provides an overview of some previously conducted studies which have been reviewed in this study. The table contains essential information such as the authors, the models used, and the evaluation criteria applied on each model. In section 2, the study addresses the origin of the CLASH database and the challenges of gathering the data, mentioning the details of the different group of parameters in the database. Also, some of the related works of the database are reviewed with clear description on the type of waves and structures, and the range of validity. Section 3 presents a brief representation on the application of machine learning, focusing on the model structure and essential requirements in machine learning. Furthermore, the section covers the gaps and issues related to machine learning in other coastal processes and addresses the related works in wave overtopping with existing challenges. Finally, Section 4 presents a summary of the study with some significant considerations for the future studies.

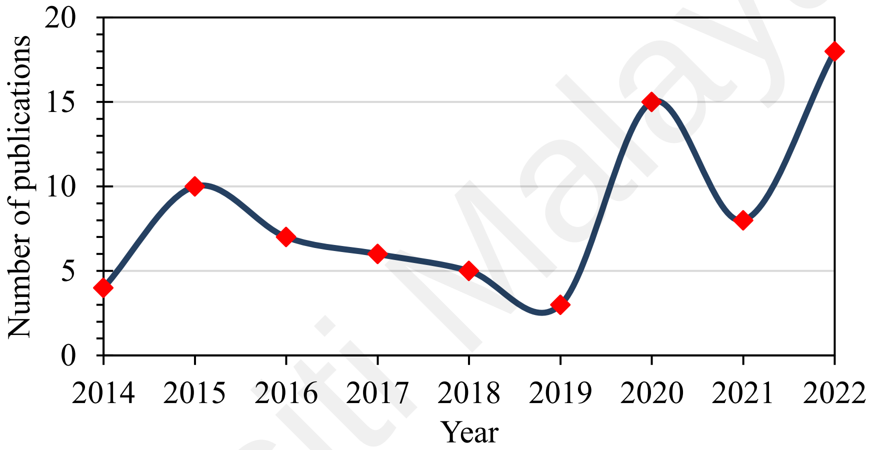


Figure 2.2: Numbers of papers published in year 2014 to 2022

Table 2.1: Previous studies including authors, methods and evaluation criteria

Authors	Scope of study	Structure type	Methods	Evaluation criteria
(Koosheh et al., 2022)	Wave overtopping at rubble mound seawalls	Rubble mound seawalls	Physical model	RMSE, BIAS
(Hosseinzadeh et al., 2021)	Wave overtopping at simple slope	Simple sloped breakwater	ANN, SVM, SVR	RMSE, BIAS, R. R^2
(den bieman et al., 2021)	Wave overtopping	Rock structure	NN, XGBoost	RMSE
(Shaeri & etemad-shahidi, 2021)	Wave overtopping	Vertical and smooth structure	ANN	RMSE, BIAS, R, DR_i
(M. S. I. Ibrahim & baldock, 2021)	Wave-by- wave overtopping	Truncated Plane Beach	Physical mode, SWASH	RMSE, NRMSE
(Den bieman et al., 2020)	Wave overtopping discharge	Simple structure	NN, XGBoost	RMSE
(K. K. Pillai et al., 2020)	Wave reflection	Berm breakwater	Physical model	RMSE, DR, BIAS
(Koosheh et al., 2020)	Wave overtopping	Armored sloped structure	Empirical formulae	RMSE, BIAS,
(Salauddin & pearson, 2020)	Wave by wave overtopping volume	Sloping structure	Physical model	RMSE, BIAS
(M. S. I. Ibrahim & baldock, 2020)	Overtopping on plane beaches	Plane beaches	Physical model	RMSE, NRMSE
(K. Pillai et al., 2019)	Wave run-up	Bermed coastal structures	Physical model	RMSE, BIAS, DR

Table 2.1: Continued

(Williams et al., 2019)	Wave overtopping	Smooth sloped and vertical structures	Physical model	R
(Formentin et al., 2017)	Wave reflection, overtopping	Smooth berms	ANN	RMSE, WI, R^2, large errors
(Zanuttigh et al., 2017)	Wave overtopping	Rubble mound slope	ANN	RMSE, WI, R^2
(Etemad-Shahidi et al., 2016)	Wave overtopping	Vertical structure	ANN	RMSE, BIAS, R, DR_i
(Zanuttigh et al., 2016a)	Wave overtopping	Coastal and harbor structure	ANN	RMSE, WI, R^2, large errors
(Zanuttigh et al., 2013)	Wave reflection	Coastal and harbor structure	ANN	RMSE, WI, R^2
(Geeraerts et al., 2007)	Wave overtopping	different Coastal structures	NN	R
(Van gent et al., 2007)	Wave overtopping	Wide range of Coastal structures	NN	RMSE

2.2 Overtopping Phenomenon

In the initial decades, overtopping simulations in laboratories exclusively utilized regular waves. However, as time progressed, the standard shifted to incorporating irregular wave generation, resulting in enhanced accuracy for the developed prediction methods. Notably, the first well-known overtopping model based on irregular wave experiments in the laboratory is Owen’s formula (Owen, 1981). Remarkably, even today, Owen's formula, derived from laboratory experiments with irregular waves, remains a standard tool for designing sloping structures (Losada et al., 2016).

The increasing impact of climate change, marked by rising sea levels and shifts in wave patterns, has heightened the significance of coastal protection structures. Ensuring the accurate prediction of wave overtopping responses in these structures is crucial for engineers striving to develop cost-effective and secure designs. The phenomenon of wave overtopping at coastal protection structures is intricate, influenced by numerous factors such as the nearshore wave climate, as well as the structure geometry and materials (Eurotop, 2018).

To gauge the extent of wave overtopping in these structures, various overtopping parameters can be taken into account, tailored to the specific structure type and project requirements. Traditionally, the mean overtopping discharge has been regarded as the main parameter for describing overtopping, serving as the primary criterion for the geometric design of structures. However, insights from experimental and field observations reveal that the maximum wave overtopping discharge during an overtopping event can be orders of magnitude larger than the mean overtopping discharge (J. W. van der Meer, 1998).

However, research continues to push the boundaries, exploring more sophisticated models and incorporating additional factors influencing overtopping. These include structure geometry, wave characteristics, and the presence of obstacles or vegetation (van Gent et al., 2007). Engineers and coastal managers acknowledge that coastal defenses minimize the risk of wave overtopping, but it takes a comprehensive understanding to recognize that seawall does not always prevent, but rather decrease overtopping.

2.3 The CLASH overtopping database

2.3.1 Establishment of database

Physical modeling, numerical modeling, or a combination of both methods can be utilized in the design of sea defenses, through the determination of allowable overtopping discharge. These models must be calibrated, and their results must be compared to prototype results (Williams et al., 2019). CLASH database is originally provided by De Rouck et al., (2005), and funded by European funding called 'Crest Level Assessment of coastal Structures by full scale monitoring, neural network prediction and Hazard analysis on permissible wave overtopping'.

Initially, CLASH database consisted of 10,532 tests gathered from universities and institutions around Europe in 2003 and 2004 (Verhaeghe, 2005). Some of the problems prior to the CLASH project was that small scale models underestimated the wave run up on a rubble mound breakwater by approximately 20% against the full scale models (De Rouck et al., 2001). Since wave run up is directly related to wave overtopping, similar argument can be applied to small scale wave overtopping tests as well (K. Pillai et al., 2019). Moreover, there is a lack of an accurate and robust wave overtopping prediction approach for all types of coastal structures. Despite vast amounts of available data on wave overtopping, these data have yet to be incorporated into a single, generic design process (Chini & Stansby, 2012). Due to these issues and findings, CLASH project was initiated to address the issue of probable scale/model impacts on wave overtopping and to develop a generic prediction approach for wave overtopping. In order to make the database more homogeneous, white spots which can also be called gaps in the database were detected and extra tests were carried out (De Rouck et al., 2009). As a result, , the goal for this project was achieved and the objectives were accomplished by having the capability to predict the allowable amount of wave overtopping (K. Pillai et al., 2017).

However, there is no thorough comparison of proposed equations for estimating overtopping rates at rubble mound sloped structures (Koosheh et al., 2020).

Due to the existing errors, Eurotop, (2018) has proposed new database consist of 17,742 tests including the original CLASH database with addition of new parameters to the database. The screening process began with the collection of original data, which was subsequently analyzed in a variety of methods, and concluded with the final database. Due to the need to maintain the confidentiality of a number of tests, it was essentially forbidden to disclose both the original data and screening approach. This means that the original dataset and the decisions made during the screening process are only known to the authors but cannot be released publicly (Steendam et al., 2005). To gather reliable overtopping data, certain concerns need to be addressed. For example, wave characteristics and types (regular or irregular), test facility (2D or 3D), and model scale (Salauddin & Pearson, 2020).

2.3.2 Challenges and improvement of the overtopping database

In order to develop a prediction technique by an artificial neural using overtopping database, each test has to be defined by a small number of parameters that summarize the test's most crucial information (J. W. van der Meer et al., 2009). The most difficult element was the determination of adequate parameters to provide a comprehensive perspective of the overtopping test and to specify the overtopping tests and cross sections of the studied structures (Etemad-Shahidi et al., 2016). 32 parameters were set to define every test (Geeraerts et al., 2007) and were divided into 3 groups: general parameters, hydraulic parameters and structural parameters (van Gent et al., 2007). By expanding the dataset of 16,165 tests on wave reflection, transmission, and overtopping compiled by Zanuttigh et al., (2013), a highly homogenous dataset of 17,942 tests was established. Several modifications were made to the CLASH dataset, including the addition of new parameters as see in Figure 2.3, and the establishment of new calculating processes (Formentin et al., 2017; Zanuttigh et al., 2017). For example, a number of new parameters have been added to the database, one of which is the diameter D (Zanuttigh et al., 2017). This parameter represents the average size of the structural elements in the run-up/down area. Generally, it provides an indication of the size of structural components, especially in the vicinity of the water level (Koosheh et al., 2020).

Further analysis of both database is presented in Table 2.2 between the parameters in the original CLASH dataset (OC) and the new CLASH dataset (NC) with defining each parameter (Eurotop, 2018). Furthermore, a new label was added to the database to point out tests with unusual characteristics, such as w for wind, p for prototype, c for current, b for bull nose, and pc for perforated caisson (Formentin et al., 2017). Despite the fact that these tests could be quite reliable, they were commonly assigned a reliability factor of 4 in CLASH, indicating a low reliability factor (Zanuttigh et al., 2017).

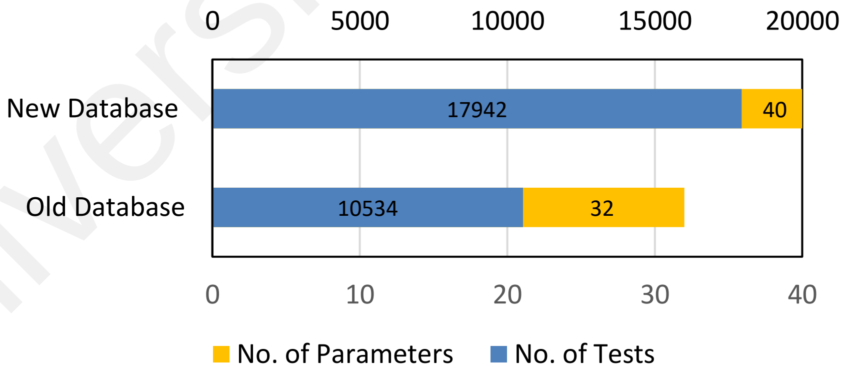


Figure 2.3: New CLASH database and Old CLASH database

Table 2.2: Parameters of original CLASH database and new CLASH database.

#	Parameter	Unit	Definition of the parameter	NC	OC	Type
1	$H_{m0\ deep}$	m	Off-shore significant wave-height	✓	✓	hydraulic
2	$T_{m\ deep}$	s	Off-shore average wave period	✓	✓	hydraulic
3	$T_{m-1,\ deep}$	s	Off-shore spectral wave period	✓	✓	hydraulic
4	$T_{p\ deep}$	s	Off-shore peak wave period	✓	✓	hydraulic
5	$h_{\ deep}$	m	Offshore water depth	✓	✓	structural
6	h	m	Water depth at the structure toe	✓	✓	structural
7	β	°	Wave obliquity	✓	✓	hydraulic
8	m	—	Foreshore slope	✓	✓	structural
9	$H_{m0\ t}$	m	Significant wave-height at the structure toe	✓	✓	hydraulic
10	$T_{m\ t}$	s	Average wave period at the structure toe	✓	✓	hydraulic
11	$T_{m-1,\ t}$	s	Spectral wave period at the structure toe	✓	✓	hydraulic
12	$T_{p\ t}$	s	Peak wave period at the structure	✓	✓	hydraulic
13	B_t	m	Toe width	✓	✓	structural
14	h_t	m	water depth on the toe of a structure	✓	✓	structural
15	$\cot\alpha_u$	—	Cotangent of the angle that the part of the structure below/above the berm makes with a horizontal	✓	✓	structural
16	$\cot\alpha_d$	—		✓	✓	structural
17	$\cot\alpha_{incl}$	—	Cotangent of the mean angle that the structure makes with a horizontal, including/excluding the berm, in the run-up/run-down zone	✓	✓	structural
18	$\cot\alpha_{excl}$	—		✓	✓	structural
19	γ_{fd}	—	Roughness factor for $\cot\alpha_d$	✓		structural
20	γ_{fu}	—	Roughness factor for $\cot\alpha_u$	✓		structural
21	γ_f	—	Roughness factor [average in the run-up/down area in the new database	✓	✓	structural

Table 2.2: Continued

22	Type	—	Type of structure and armor unit	✓	✓	structural
23	S	—	Spreading factor	✓		structural
24	G_c	m	Crest width	✓	✓	structural
25	h_b	m	Berm submergence	✓	✓	structural
26	B_h	m	Horizontal berm width	✓	✓	structural
27	A_c	m	Wall height with respect to SWL	✓	✓	structural
28	D	m	Average size of the structure elements in the run-up/down area	✓		structural
29	D_u	—	Size of the structure elements along $\cot\alpha_u$	✓		structural
30	D_d	—	Size of the structure elements along $\cot\alpha_d$	✓		structural
31	B	m	Berm width	✓	✓	structural
32	P_{ow}	—	Percentage of the waves resulting in overtopping = $(N_{ow} / N_w).100$	✓	✓	hydraulic
33	$\tan\alpha_B$	—	Berm slope	✓	✓	structural
34	R_c	m	Crest height with respect to SWL	✓	✓	structural
35	CF	—	'Complexity Factor' The complexity of the test is indicated by a score with a possible range of 1 to 4.	✓	✓	general
36	K_t	—	(bulk) wave transmission coefficient	✓		hydraulic
37	K_r	—	(bulk) wave reflection coefficient	✓		hydraulic
38	RF	—	'Reliability Factor' The reliability of the test is indicated by a score with a possible range of 1 to 4.	✓	✓	general
39	q	$m^3/s/m$	Average specific wave overtopping discharge	✓	✓	hydraulic
40	Core data	—	Flag indicating the inclusion/exclusion from the core data of the ANN training	✓		general

A reasonable amount of gathered information for all test series must be obtained for a comprehensive and trustworthy overtopping dataset. Information about wave properties, test structure and corresponding overtopping discharge is needed. However, information regarding the facility utilized to conduct the tests, the measurement process, and the accuracy of the job accomplished were all gathered (Verhaeghe, 2005). In addition, to avoid any issues in the results, reliability factor (RF) and complexity factor (CF) were defined for each test in pre-processing phase of the development of the generic model (van Gent et al., 2007). Some of the parameters and tests will need to be filtered again cause fewer parameters as input data is required for the prediction approach (van Dongeren et al., 2018). These parameters are given in the database to give as much information as possible about each test; nevertheless, not all of it will be utilized to develop the model (Zanuttigh et al., 2016a).

In brief, the final ‘expanded’ database has additional 8 parameters compared to the original CLASH database, bringing the total number of parameters to 14 hydraulic parameters, 23 structural parameters and 4 general parameters (Formentin et al., 2017). By using the new database and applying the new suggested formula, with range of validity $0.05 < R_c / H_{m0} < 0.08$ Gallach-Sánchez et al., (2021) found that it has improved the accuracy of the overtopping prediction, where the value of the RMSE is 21% and, better than the prior prediction using the CLASH database (J. van der Meer & Bruce, 2014). Also, it was noticed that the new suggested formula improves the prediction of overtopping rate for steep low crested structures by increasing the accuracy for zero freeboards (R_c=0) with 35% reduction of RMSE and extremely small relative crest freeboards ($0.11 > R_c$/Hm0 > 0) (16% reduction of RMSE), as well as for very steep slopes ($0.27 > \cot \beta > 0$) with 31% reduction in RMSE and vertical structures ($\cot \beta = 0$) with 24% reduction in RMSE (Gallach-Sánchez et al., 2021).

2.3.3 Empirical models and overtopping rate formulae

J. De Rouck et al., (2005) was the first to study the impact of model scale on wave overtopping discharge measurement for several types of structures including rubble mound structures and vertical walls. They found excellent correlation between the field prediction and the laboratory models for vertical walls, with few variations explained by model impacts. However, for rubble mound structures, a clear difference was noticed particularly in cases with small overtopping values. As the slope gets longer and flatter, the contrast between two elements becomes more noticeable (Koosheh et al., 2021). Also, for rubble mound sea walls, limited records were available in the overtopping database (around 120 data) (Koosheh et al., 2022). Figure 2.4 compares wave overtopping formula for various type of structures. Limited number of tests is available for steeper slopes while a significant number of data is accessible to structure with slopes of tan $\alpha = 0.5$ (Hosseinzadeh et al., 2021). This describes the existing gaps in the database related to the rubble mound sea walls and show the improvement needed for the key parameters of the database (Etemad-Shahidi et al., 2021). The expected behavior of the structure is more or less given by the recession, which can range from minimal to complete restructuring, depending on the classification. This impacts the selection of berm width (W. Chen et al., 2020). Clearly, the smaller the stability number of the berm rock, the more stable and recession-resistant the construction will be. However, minimizing recession should not be the exclusive purpose of berm width design. The breadth of the berm should be much greater than the anticipated recession. With a lesser recession, the capacity to withstand brutal environment increases. The capability of a structure to resist extreme conditions is knows as its resiliency, and this resiliency should play a part in designing the berm width, but it has never been officially stated in design guidelines (J. W. van der Meer, 2017).

General form of empirical overtopping formula (Eurotop, 2018):

$$\frac{q}{\sqrt{g.H_{m0}^{3}}} = \frac{0.023}{\sqrt{tan\alpha}}\Upsilon_{b}.\,\varepsilon_{m-1,0}.\,\mathbf{exp}\left[-\left(2.7\frac{R_c}{\varepsilon_{m-1,0}.H_{m0}.\Upsilon_{b}.\Upsilon_{f}.\Upsilon_{\text{ß}}.\Upsilon_{v}}\right)^{1.3}\right] \tag{2.1}$$

With a maximum of

$$\frac{q}{\sqrt{g.H_{m0}^{3}}} = 0.09.\,\mathbf{exp}\left[-\left(1.5\frac{R_c}{H_{m0}.\Upsilon_{f}.\Upsilon_{\text{ß}}.\Upsilon}\right)^{1.3}\right] \tag{2.2}$$

Where γ_b is the influence factor for a berm, γ_f is the influence factor for roughness elements on a slope, γ_β is the influence factor for oblique wave attack, and γ_v is the influence factor for a wall at the end of a slope.

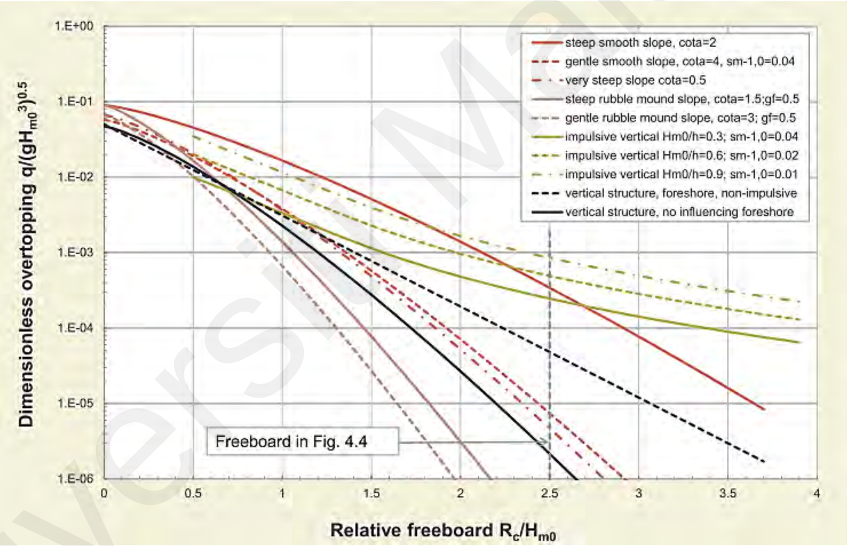


Figure 2.4: Comparison of wave overtopping formulae for different types of structures. (Eurotop, 2018)

Due to the limitations in the field measurements of wave overtopping and the high cost/complications of field studies in terms of installation and maintenance, most of the existing tests are being conducted under laboratory environment (Koosheh et al., 2021). Consequently, the majority of empirical formulations have been developed based on laboratory results. However, selecting a suitable modelling scale and technique is challenging in data collection process for subsequent study (Shaeri & Etemad-Shahidi, 2021). There are numerous sources of certainty regarding the wave overtopping processes, that led to various approaches in overtopping measurements. Koosheh et al., (2022) has conducted a series of overtopping experiments to address the gap in the CLASH database. The experiment is conducted in a flume with a 0.8 m depth, length and width of 22.5 m and 0.5 m respectively. Also, it is equipped with a piston type wave maker which can produce regular waves and irregular waves. Three capacitance wave gauges were installed at the structure's toe to monitor the water's free surface and estimate wave parameters. Based on this experiment, it was shown that the proposed formula by Koosheh et al., (2022) has RMSE value of 0.51 m^3/s/m for prediction of wave overtopping discharge for long waves which represents about 40% improvement in production accuracy compared to the formulas by Owen, (1981) and Eurotop, (2018) as can be seen in Figure 2.5. Furthermore, 50% of test data (obtained from ETS) is used to derive the formula by Eurotop, (2018), whereas the test dataset is completely unseen for the proposed formula by Koosheh et al., (2022).

Many existing laboratory approaches are based on the evaluation of the temporal evolution of the overtopped water volume stored in a container at the end of the structure. The volume of the container is then estimated by either measuring the water mass or water level (W. Chen et al., 2020). There are two types of overtopping, i.e., run up and over the face of the structure in coherent water mass, and spray overtopping, which typically occurs when waves break seaward of the structure (Koosheh et al., 2021). When there is no substantial wind velocity, the contribution of the second type to the overtopping volume may be insignificant (Eurotop, 2018). However, by applying the formula from Eurotop, (2018), the results of the empirical predictions were not satisfactory with 72% of the data lying within the prediction ranges. For prediction of mean overtopping rates, the empirical predictions of Etemad-Shahidi & Jafari, (2014) and Eurotop, (2018) were adopted and it was found that both models provided good estimation of wave overtopping. However, according to Salauddin & Pearson, (2020), it is also notable that the values obtained show a weak correlation with the predicted values of Eurotop, (2018) than those found by Etemad-Shahidi & Jafari, (2014).

Table 2.3: List of empirical overtopping formulae for CF = 1 & 2

Authors	Overtopping formula	Range of validity
(Owen, 1981)	$\frac{q}{gH_{m0}T_m} = a\exp\left(-b\frac{R_c}{H_{m0}}\sqrt{\frac{s_{om}}{2\pi}}\frac{1}{\gamma_f}\right)$	$0.03 \leq S_{m-1,0} \leq 0.07$
(Goda, 2009)	$\frac{q}{gH_{s,toe}^3} = \exp\left(-A - B\frac{R_c}{H_{s,toe}}\right)$	0 ≤ cot∞ ≤ 7
(Etemad-Shahidi & Jafari, 2014)	$\frac{q}{\sqrt{gH_{m0}^3}} = 0.032.\exp\left[-2.6\left(\frac{R_c}{H_{m0}}\right)^{1.6}.\left(\varepsilon_{m-1,0}\right)^{-1.26}\right]$	$\frac{R_c}{H_{m0}} \leq 1.62$
	$\frac{q}{\sqrt{gH_{m0}^3}} = 0.032.\exp\left[-5.63\left(\varepsilon_{m-1,0}\right)^{-1.26}.-3.283\left(\frac{R_c}{H_{m0}} - 1.62\right)^{0.83}\right]$	$\frac{R_c}{H_{m0}} > 1.62$
(Eurotop, 2016)	$\frac{q}{\sqrt{gH^3}} = \frac{0.023}{\sqrt{tan\beta}}\varepsilon_{m-1,0}exp\left[-\left(2.7\frac{z_c}{\varepsilon_{m-1,0}H_{m0}}\right)^{1.3}\right]$	$0.01 \leq S_{m-1,0} \leq 0.04$ $Ir_{m-1,0} = 1.8$
(Eurotop, 2018)	$\frac{q}{\sqrt{gH_{m0}^3}} = \frac{0.023}{\sqrt{tan\alpha}}.\varepsilon_{m-1,0}.\exp\left[-\left(2.7\frac{R_c}{\varepsilon_{m-1,0}.H_{m0.\gamma_f}}\right)^{1.3}\right]$	$Ir_{m-1,0} < 2$
	$\frac{q}{\sqrt{g.H_{m0}^3}} = 0.09\exp\left[-\left(1.5\frac{R_c}{H_{m0\gamma_{f\cdot\gamma_\beta}}}\right)^{1.3}\right]$	$0.04 \leq S_{m-1,0} \leq 0.06$ $2.5 < Ir_{m-1,0} < 4$

Table 2.3: Continued

(M. S. I. Ibrahim & Baldock, 2020)	$V^* = a^* \frac{HL_0}{2\pi} tan\beta \frac{(R - z_C)^2}{R^2}$	0.01 ≤ $S_{m-1,0}$ ≤ 0.04 0.40 ≤ $Ir_{m-1,0}$ ≤ 1.88 a* = 0.313
	$V^* = a^* \frac{HL_0}{2\pi} \sqrt{tan\beta} \frac{(R-z_C)^2}{R^2}$	0.01 ≤ $S_{m-1,0}$ ≤ 0.04 0.40 ≤ $Ir_{m-1,0}$ ≤ 1.88 a* = 0.124

(Gallach-Sánchez et al., 2021)	$\frac{q}{\sqrt{gH_{m0}^3}} = a.\exp\left(-\left(b\frac{R_c}{H_{m0}.\Upsilon_f.\Upsilon_ß}\right)^c\right)$	R_c/ H_{m0} > 0.8
		a = 0.109 – 0.035.(1.5 – cotα) for cotα ≤ 1.5 and a = 0.109 for cot α > 1.5
		b = 2 + 0.56 (1.5-cotα)$^{1.3}$ for cotα ≤ 1.5 and b = 2 for cot α > 1.5
		c = 1.1

(Koosheh et al., 2022)	$0.034\ \exp\left[-4.97\left(\frac{R_c}{H_{m0\gamma_f}}\right)^{1.12}(S_m - 1,0)^{0.35}\right]$	0.018 ≤ $S_{m-1,0}$ ≤ 0.057 2.80 ≤ $Ir_{m-1,0}$ ≤ 5.03

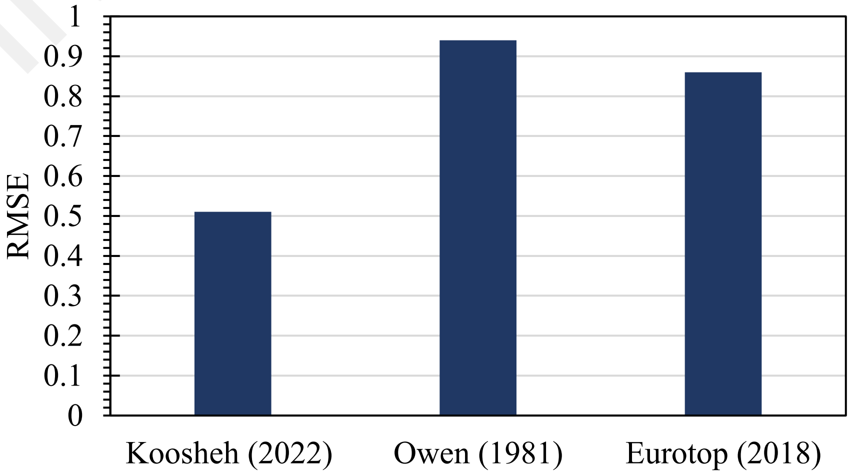


Figure 2.5: Comparison between RMSE results for different overtopping formulae. (Koosheh et al., 2022)

M. S. I. Ibrahim & Baldock, (2020) used the fundamental methodology as Baldock et al., (2012) and Baldock et al., (2005) to conduct 323 overtopping tests comprising of random and monochromatic waves. The bathymetry consisted of 8.5m long horizontal section and the beach gradients were 1V:10H and 1V:5H. The width of the tank used to measure the overtopping was 0.72m and 0.45 m and 0.16 in length and depth respectively. According to M. S. I. Ibrahim & Baldock, (2020), both the empirical model of the Eurotop, (2016) model and the theoretical model of Peregrine & Williams, (2001) can be formulated with consistent coefficients incorporating the positive volume flux term. Semi-empirical equations incorporating new scaling law were derived from the formula of the Eurotop, (2018) to calculate the overtopping volume. Theoretically, the new scaling law treated the overtopping volume via the positive volume flux, a parameter that corresponds to the positive volume due to the displacement/movement by the wavemaker. Although the results of the Eurotop, (2016) is slightly better, it still proves that the existing overtopping formula can be converted to new scaling laws with minimal RMSE value. Another experiment is performed by Williams et al., (2019) in a wave flume of 15 m and 0.23 m in length and width respectively, while the depth is 0.22 m. The bathymetry consists of an impermeable smooth slope with a gradient of 1V:2.55H and a vertical wall with freeboard of $R_c = 0.06$m. Utilizing a peak detection method, each overtopping event was recorded and the difference between following peaks was detected to calculate the overtopping volume. According to Williams et al., (2019), two limitations in this methodology have been notified. First, there was some residual uncertainty in the predicted volume rates due to the noise in the date, which was created by oscillations in the surface of the water. Another limitation is that in relatively high overtopping conditions, where several overtopping events occurred in rapid succession, it is likely that minor overtopping events were not recognized in the signal and were, thus, lost.

On the other hand, the experiment done by Salauddin & Pearson, (2020) showed results that were identical to actual values as the difference between total measured volumes and actual volumes was approximately 0.7% and the RMSE recorded was 9.38 ml. Furthermore, the RMSE values in Figure 2.6 showed that the projection formula of Eurotop, (2018) outperforms those of Etemad-Shahidi & Jafari, (2014) and Goda, (2009) for estimation of overtopping rates at a sloping wall with an impermeable foreshore. The experiment was conducted in accordance with Eurotop, (2018) requirements for typical two-dimensional wave flume investigations in a wave channel with the studies of 22 m in length, 0.6 m width and 0.7 m in depth. Despite the fact that this experiment produced positive results, it was noticed that the load-cell sensitivity was restricted to 5-9 ml of overtopping volume, therefore lower overtopping volumes resulted in a bigger measurement error than large values (Salauddin & Pearson, 2020).

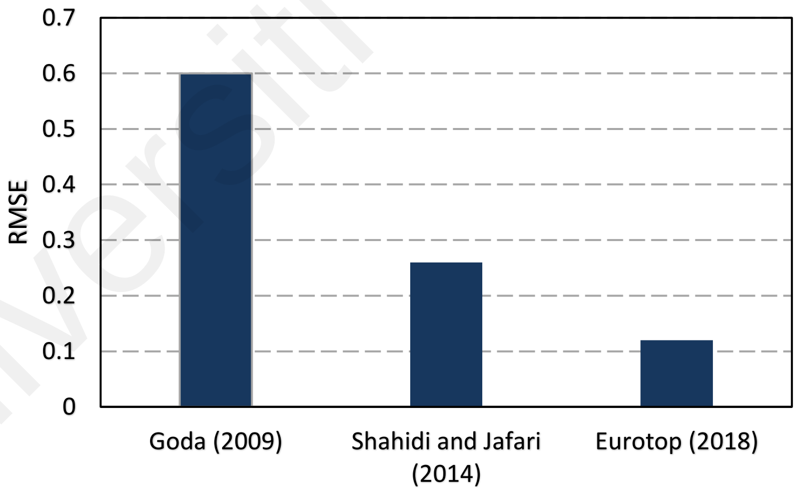


Figure 2.6: RMSE results for cases with sloping wall and impermeable foreshore

2.4 Application of Machine Learning in wave overtopping

2.4.1 Model structures/requirements in Machine Learning

Recently a huge amount of data can be retrieved openly from various sources and modern technology. Such data commonly contain a lot of useful information that can be used in different aspects. For example, in coastal engineering, a lot of prediction can be done using information such as the wave overtopping and the wave runup for individual waves (Rogers, 2020). However, extracting the information needed from the data is not an easy task since most of the databases include thousands of tests and parameters. For that reason, machine learning is needed to provide reliable and effective techniques to adapt and use this data in a functional way. Artificial neural networks (ANNs), support vector machine (SVM), and hybrid systems adaptive neuro-fuzzy inference system (ANFIS) are some well-known AI techniques and have been using by a lot of researchers in general and in the field of civil engineering specifically. Recently, Extreme Gradient Boosting (XGB) has been used in a few studies that related to wave overtopping. As illustrated in Figure 2.7, it can be seen that 135 papers are using ANN, while 43, 57 and 2 articles are applying ANFIS, SVM and XGB respectively. Due to their accuracy in fitting a very small collection of data and their modest development, ANN and ANFIS tend to be among the most popular AI approaches (Muslim et al., 2020).

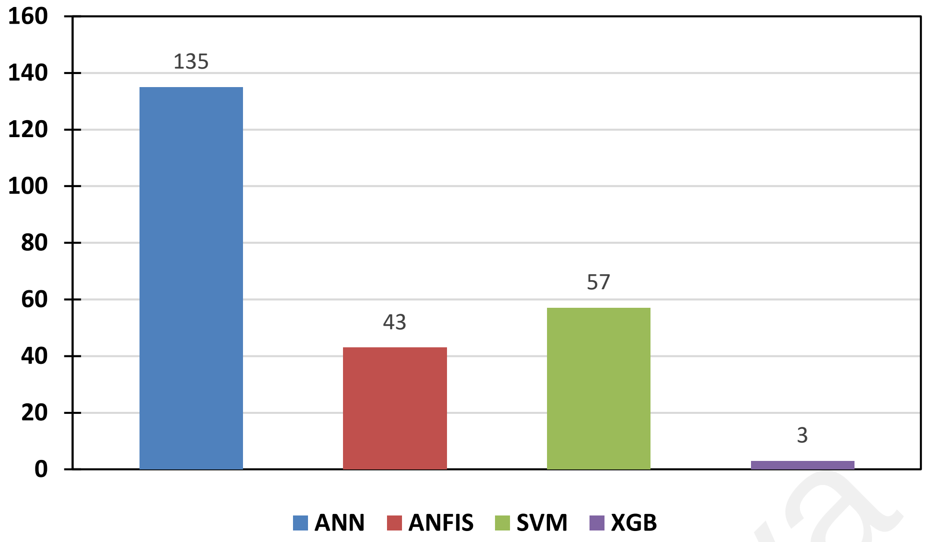


Figure 2.7: Number of papers for selected AI models related to wave overtopping works

2.4.1.1 Artificial Neural Network (ANN)

The neural network (NN) is a mathematical model that simulates neuron behavior in humans. In 1943, McCulloch and Pitts were the first to present the foundations of NN. Figure 2.8 depicts the model fundamental structure, which consists of input layer, output layer and one connection weight layer where X_1, X_2,….X_n is the input neuron, W_1, W_2,….W_n is the connection weights, S is the total weighted input signals, f(s) is the activation function and y is the output (Wee et al., 2021). Improvements to the neural network model have been made throughout time to produce a robust prediction model (Jumin et al., 2020). The AI models have demonstrated satisfactory performance with up to 90% accuracy. The obvious differences, however, are in the input parameters and time measuring used in the models (Afan et al., 2016). Artificial neural networks (ANNs) are a collection of densely connected processing units that manage parallel-distributed information systems and have similar idea as biological neural networks in the human brain (Allawi et al., 2018). Daily inflow forecasting has been studied by Elizaga et al., (2014), using a back propagation technique based on an artificial neural network. Indicators such as RMSE, MAE and correlation coefficient are examples of the statistical metrics which have been used to assess the performance of this method. As shown in Figure 2.9, the findings yielded that the ANN technique anticipated the most accurate inflow, with excellent correlations between current and projected values and minimal errors with RMSE value of 0.020 m^3/s/m. Filippo et al., (2012) applied the ANN method to conduct sea level forecasting work. They found that the ANN performed well in predicting meteorological tides utilizing wind, air pressure, and estimated data of harmonic tide. However, Igboanugo, (2013) tested five different ANN models, each with one to five inputs. Although the model accuracy originally increased in direct proportion to the number of inputs, the accuracy began to fall once the fourth input was added. This indicates the limitation of the model against higher order of input data as described by Wee et al., (2021).

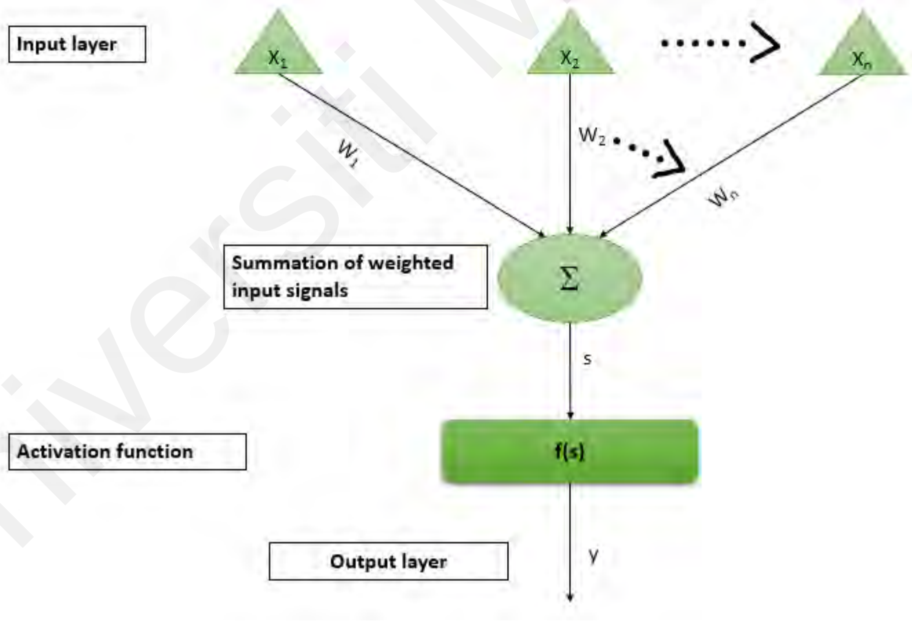


Figure 2.8: Structure of neural network model

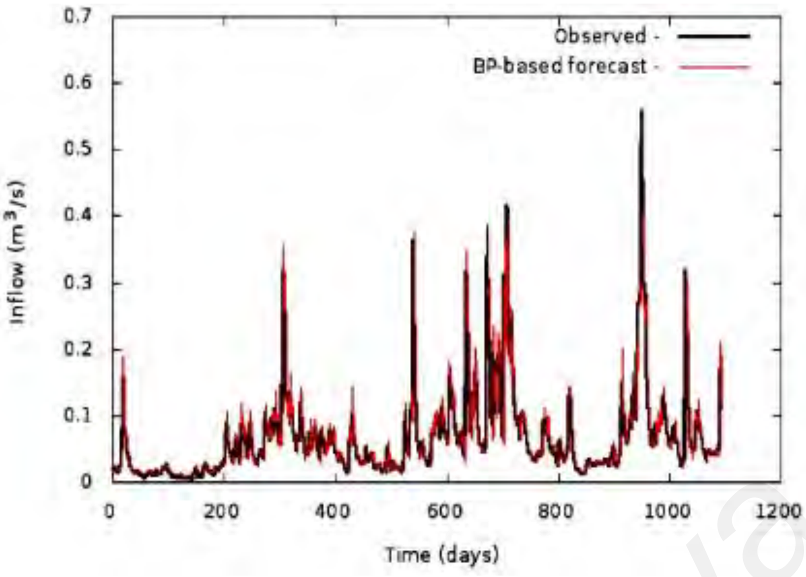


Figure 2.9: Observed average inflow vs backpropagation forecasts for validation (Elizaga et al., 2014)

2.4.1.2 Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS is an AI technique with a flexible statistical framework that may identify complicated non-linearity and difficulties caused by randomness and inaccuracy among variables without attempting to infer the nature of trends (Hanoon et al., 2021). Initially, ANFIS is based on two theories, i.e., The Generalized Neural Network and fuzzy inference system. The Generalized Neural Network, a multilayer feed-forward neural network made up of nodes and neurons that connect the input to the output across numerous layers, and the fuzzy system, a neural network theory and reasoning system, make up the first theory. While the second theory is referred to as the 'fuzzy inference system' and it is essentially stated as if-then set of rules, with each fuzzy rule acting as a local characterization of network behaviors ((K. S. M. H. Ibrahim et al., 2022). As shown in Figure 2.10, ANFIS model consists of 5 layers and it adopts a unique algorithm i.e., a hybrid learning algorithm which is divided into gradient descent and least squares approaches, to update the variables on a regular basis in the equations (Muslim et al., 2020). ANFIS has been used in various projects related to water and coastal engineering and each project shows the potential and the accuracy of ANFIS compared to other models. Ahmad et al., (2016) predicted the daily evaporation from the reservoir using the RBF-NN and ANFIS model. In comparison to the ANFIS model, the findings proved the resilience of the RBF-NN model as well as its capacity to achieve high accuracy. While the presented models were a powerful tool for fitting the evaporation phenomenon and gave sufficient accuracy, they lacked the capability to accurately anticipate peak values (Allawi et al., 2018). For sea level prediction in Sandakan, ANFIS has been used and the output of the ANFIS model was showing better performance than MLP-ANN with approximately 6.23% improvement. In both the training and testing stages, it was found that the ANFIS model can more accurately describe the behavior of the mean sea level pattern than MLP-ANN (see Figure 2.11). The majority of the anticipated values are close to the observed mean sea level (Muslim et al., 2020). Another study was conducted by Shafaei & Kisi, (2016) using three models which are rainfalls, the drainages of Shihmen reservoir and the Danshuei estuary tide. The results were satisfactory, and it implied that the model is successful and that it can be used in other studies. Also, Wee et al., (2021), described the good performance of ANFIS model in the prediction of water level forecasting while ANFIS models outperform ANNs in some aspect, the margins of error in the end outputs are quite negligible. Nevertheless, it is determined that ANFIS is a preferable alternative due to its superior learning ability for same complexity networks with significantly lower convergence error (Afan et al., 2016). Yet, as indicated by Ibrahim et al., (2022), a drawback of ANFIS is that as the coding becomes more complex, the occurrence of incorrect patterns also rises. Although ANFIS is capable of highly nonlinear mapping and outperforms MPL and other typical linear techniques of equivalent complexity, it frequently shows a quick convergence followed by a phase of significant instability. Adjustable parameters are still required by ANFIS, which should be expected and attained by trial and error approaches (Adnan et al., 2019).

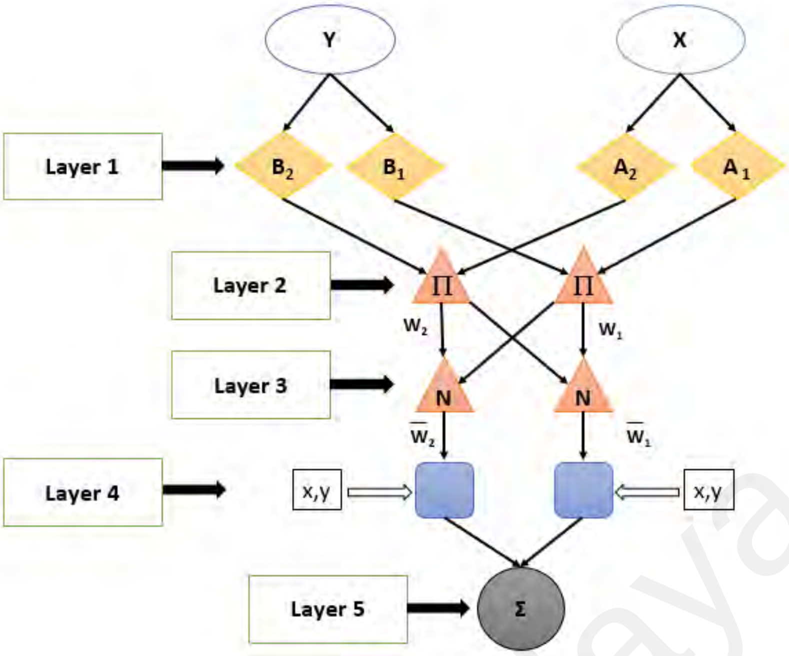


Figure 2.10: Structure of ANFIS

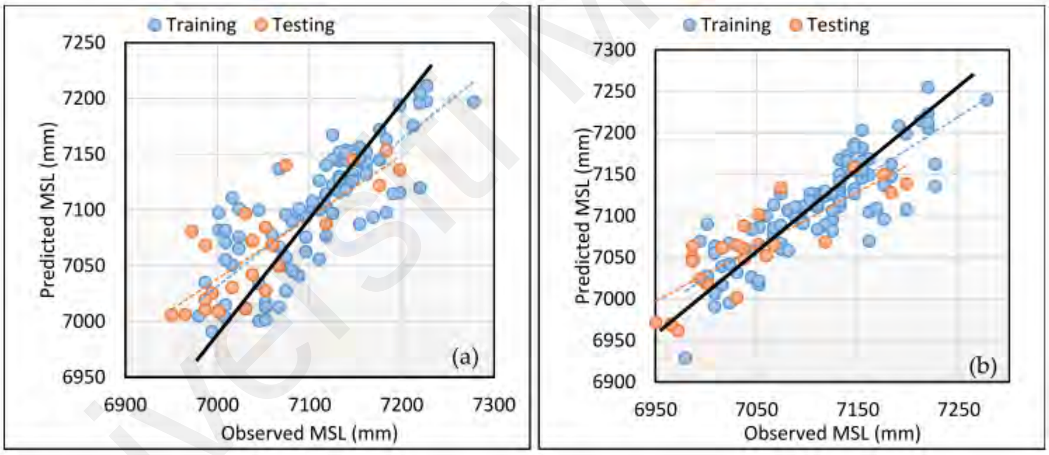


Figure 2.11: Comparison between (a) ANN and (b) ANFIS performance (Muslim et al., 2020)

2.4.1.3 Support Vector Machine (SVM)

Over the past four decades, the support vector machine (SVM) technique has gained prominence as a new type of statistical learning that has been demonstrated to be a quick and effective tool. SVM is a supervised machine learning algorithm that optimizes the difference between two groups (Ibrahim et al., 2022). It is built on the basis of locating a line, a plane, or some surface that divides two categorization groups (Afan et al., 2016).

SVM also ensures an optimal physique and zones of crucial abilities (Wee et al., 2021). Figure 2.12 depicts the SVM operation structure.

SVM models have several advantages over ANN models, such as the capacity to resolve small data in terms of nonlinear, high-dimensional, localized minimums and other partial components. Additionally, SVM has a modular design that permits independent implementation of component designs (R. K. Ibrahim et al., 2019). SVM has a lot of advantages such as the ability to avoid the problem overfitting which happens in ANN and ANFIS modelling. Also, SVM is defined as a convex optimisation problem that uses efficient approaches to solve the problem of local minima. SVMs give strong out-of-sample generalization when a proper kernel is chosen, such as the Gaussian kernel. This suggests that SVMs can be resilient even when the model sample data is biased during the training phase by the selection of appropriate generalization evaluation values (Afan et al., 2016). Furthermore, because of its optimization technique, the SVM can resolve a linearly limited quadratic programming function. Moreover, SVM might be strengthened in terms of simplicity, numerical optimization, and the training set selection procedure (Wee et al., 2021). However, deficient efficiency when applications are more than sample is one of the disadvantages about SVM (Ibrahim et al., 2022). Another limitation is that training and testing sessions are time consuming, which is inconvenient for real-world applications. Despite the fact that SVMs offer good generalization, it may exhibit slowness during the testing phase. To add on, the computational complexity and enormous memory needs of the needed quadratic programming in large-scale applications are likely the most critical drawback with SVMs from a pragmatic perspective (Afan et al., 2016).

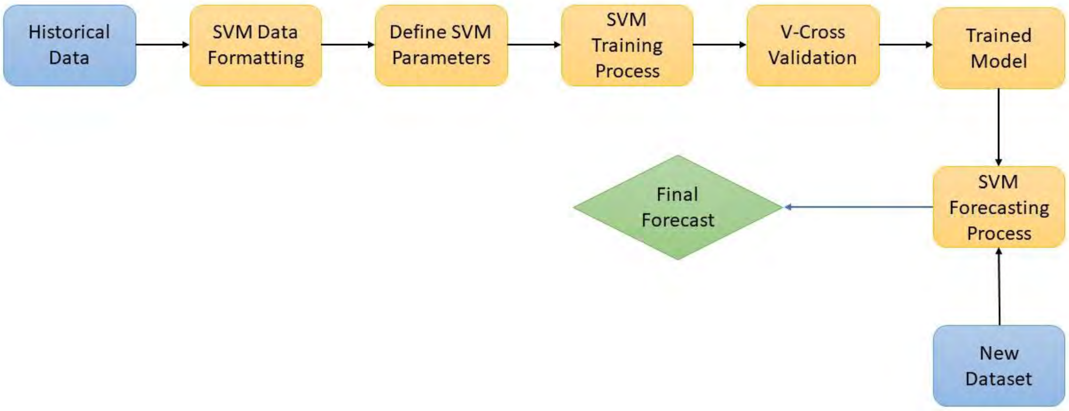


Figure 2.12: Structure of SVM operation

2.4.1.4 Extreme Gradient Boosting (XGB)

The XGBoost (Extreme Gradient Boosting) model is a powerful machine learning algorithm that has proven to be highly effective in various applications (T. Chen & Guestrin, 2016), including wave overtopping analysis. XGBoost is an implementation of gradient boosting, a tree-based ensemble method that has become popular in recent years. As shown in Figure 2.13, the algorithm builds multiple decision trees in a sequential manner and combines the results to produce a final prediction. The unique aspect of XGBoost is the use of a second-order Taylor series approximation to the loss function and a weight-based sampling method to construct a diverse set of trees. These two features, combined with the use of an optimized tree-splitting algorithm, result in faster and more accurate predictions compared to traditional gradient boosting algorithms (T. Chen & He, 2014). According to den Bieman et al., (2021), XGBoost in wave overtopping analysis has been shown to outperform other machine learning models such as Artificial Neural Networks, Adaptive Neuro-Fuzzy Inference Systems, and Support Vector Machines. Furthermore, XGBoost is highly scalable and can work with large datasets, making it a suitable choice for coastal engineering applications where large datasets are often encountered. While XGBoost has shown great potential in various applications, including wave overtopping analysis, it also has several limitations that must be considered. Firstly, XGBoost requires a significant number of computational resources and time, making it unsuitable for real-time applications where quick predictions are necessary. Additionally, XGBoost is sensitive to the choice of hyperparameters, which can greatly impact the performance of the model. This requires extensive tuning and experimentation to find the optimal hyperparameters for a given problem, which can be time-consuming and challenging. Moreover, XGBoost may overfit the data if not properly regularized, resulting in poor generalization performance. Furthermore, XGBoost may also struggle with imbalanced datasets, where one class is much more prevalent than the other, leading to biased predictions. Finally, XGBoost relies on the assumption of linear relationships between the predictors and the response variable, which may not always hold in real-world scenarios. These limitations should be carefully considered when deciding whether to use XGBoost for wave overtopping analysis or any other application. Despite its limitations, XGBoost offers several advantages that make it a popular choice for various applications, including wave overtopping analysis. Firstly, XGBoost is highly accurate, consistently outperforming other machine learning models in terms of prediction accuracy. This is due to the use of an optimized tree-splitting algorithm and the combination of multiple decision trees, which results in a more robust and generalizable model. Secondly, XGBoost is highly scalable and can handle large datasets, making it a suitable choice for applications where large amounts of data are encountered. Additionally, XGBoost is flexible and can handle both numerical and categorical data, making it a versatile choice for a wide range of problems. Furthermore, XGBoost is an open-source software and has a large community of users and developers, providing access to a wealth of resources and support. Finally, XGBoost is easy to implement and has a user-friendly interface, making it accessible to users with varying levels of technical expertise.

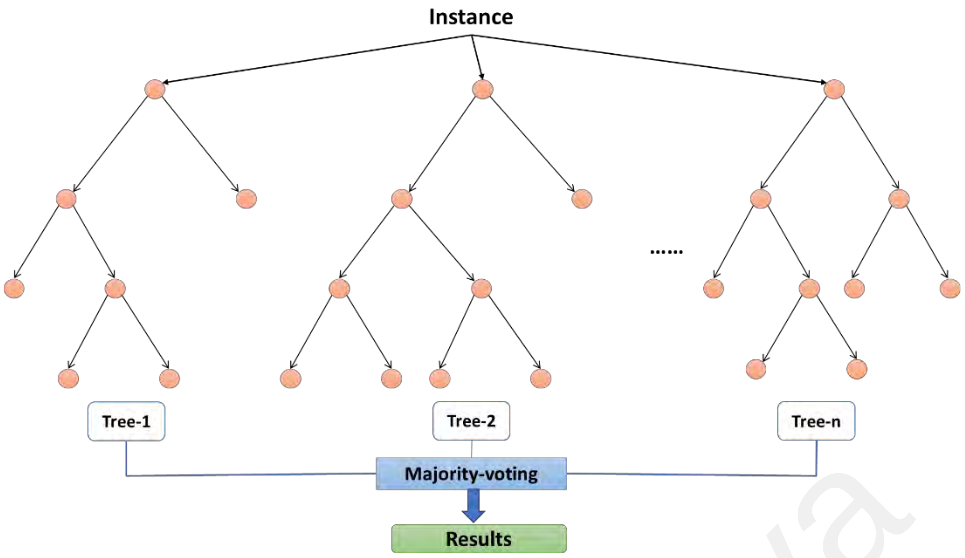


Figure 2.13: XGB Architecture

2.4.2 Applications of machine learning in coastal research

Coastal structures are designed and built to protect coastal areas from waves and storms and high-water levels during severe storms. To ensure optimum protection of the stakeholders and property on and behind the structures, overtopping rates must be smaller than the permitted rate in both normal and extraordinary operating situations (Goda, 2009). During the last decade, a lot of methods have been utilized to predict the overtopping rate. Overtopping rate equations are frequently derived using dimensional analysis and regression approaches using data from laboratory studies. However, there is a wide range of results (up to two orders of magnitude) across different methodologies and measurements, particularly for minor overtopping rates (Lykke Andersen, 2006). De Gerloni et al. (1991) investigated several vertical and composite breakwater structures utilizing a random wave flume test. They observed that the ratio of maximum to mean wave overtopping rates is affected by the geometry of the structure. Meanwhile, Van der Meer and Bruce (2014) recently reviewed the CLASH dataset to normalize relations its slope and vertical structural formulae, and discovered that the rate of overtopping is affected by relative berm depth, wave impulsiveness, and relative freeboard. Formentin et al., (2017), developed a new ANN technique for predicting the overtopping discharge for a wide variety of structure geometry and wave scenarios. To test the generalization ability of the new model, 261 tests have been excluded from the training dataset including 50 tests on Tetrapod breakwater structure, 43 tests on smooth dikes and 20 tests on smooth berms. Also, Zanuttigh et al. & Formentin, (2017) proposed a new idea to test the applicability in the absence of new experimental results. This was accomplished by removing certain separate datasets from the training database, retraining the ANN using the narrower database, and then applying the retrained ANN to predict the excluded data. As shown in Figure 2.14, around 70% of the predictions are within 95% confidence intervals, with no significant error noted. By comparing the results, it was shown that the new ANN model has more accuracy than the old existing models done by J. W. van der Meer et al., (2005). Yet, the latest ANN technique cannot be used outside of its training area. In the event of rubble mound structures and small overtopping values (less than 1 l/s per meter), the projections are influenced by model impacts due to the fact that the model was only trained on laboratory experiments. Furthermore, a correction factor (fq) is necessary for ANN predictions and the use of formulas to recalculate the values of q under prototype circumstances (Zanuttigh et al., 2017).

Despite the new ANN model, mean wave overtopping discharge was predicted using a new machine learning method called gradient boosting decision trees by (Den Bieman et al., 2020). By using the same CLASH database for training and testing, it was found that the new model is outperforming the existing NN model by a factor of 2.8 in terms of error reduction. The XGB model prediction errors defined by bootstrap resampling are thought to be very low, though not fully describing the diversity in the training data (den Bieman et al., 2021). Moreover, high performance on unknown data indicates that the model has not been overfit. Therefore, the model is not fully reliable to fit to additional data or predict future observations (Koosheh et al., 2021). On the basis of the theory that low overtopping data are more likely to be impacted by measurement errors and the presumption that (q = 10e-6 m^3/s/m) provides a sufficient value to differentiate between "negligible" and "substantial" overtopping, values of $q \leq 10^{-6}$ were removed from the training dataset for the existing ANN (van Gent et al., 2007). Data with $q < 10^{-5}$ was the only one used for training in the existing model of ANN (Zanuttigh et al., 2016b). Zanuttigh et al., (2016a) observed that the error might be because of the exclusion of values smaller than q = 10e-6 from the training database. Hosseinzadeh et al., (2021) introduced another technique which is kernel-based models that has a good prediction compared to ANN models. Furthermore, for kernel-based models, when the optimum structure modification is obtained automatically, no manual structural modification is necessary. Nevertheless, this method does not provide formulas like previous soft computing models (Zanuttigh et al., 2013).

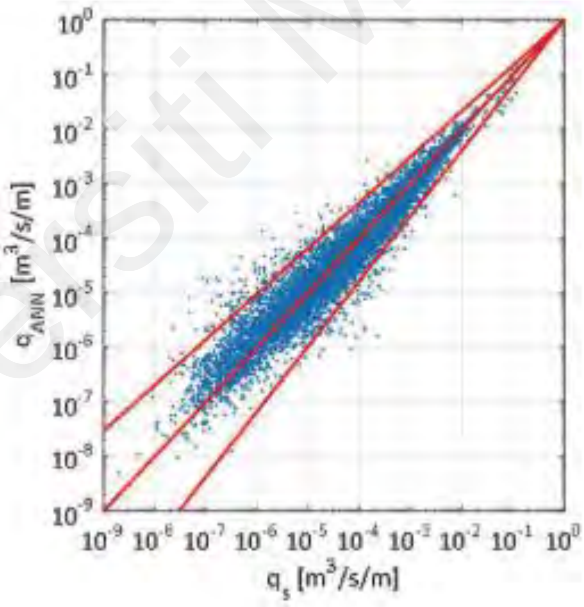


Figure 2.14: Model performance of ANN model with confidence intervals of 95% (Formentin et al., 2017)

Wave overtopping prediction was done by den Bieman et al., (2020) using the XGBoost technique. It was noticed that the NN empirical model done by both TAW, (2002) and Eurotop, (2018) show a large amount of scatters, while the XGB model exhibit a small amount of scatters (Figure 2.15). This indicates that the XGB model done by den

Bieman et al., (2021) has a small value of RMSE which shows how reliable the model is. This is because the method used was splitting the dataset between the training and testing. However, in order to acquire trustworthy numerical data, numerical models must be extensively tested and validated using physical model data.

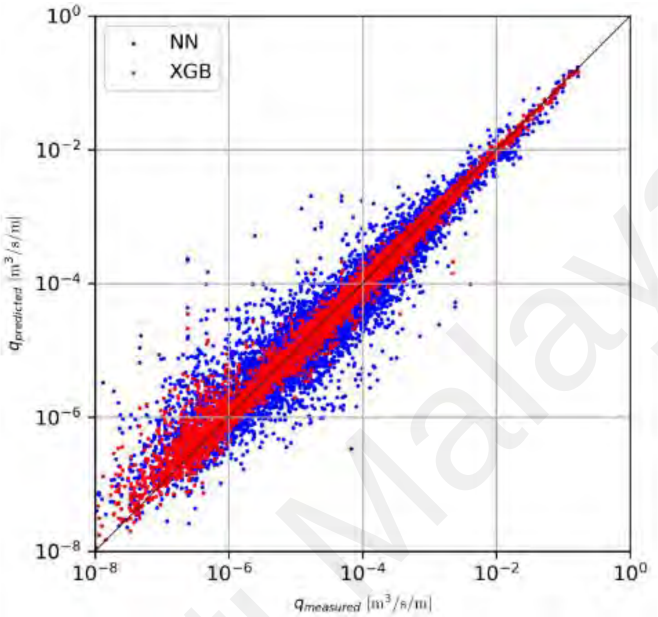


Figure 2.15: Scatters from XGB model and NN model (den Bieman et al., 2020)

2.5 Concluding remarks

Based on the assessment of previous case studies, the availability of data and the ability of the AI models to handle the real problems are crucial in selection of model. There are distinct advantages and disadvantages for each AI model that have been observed in the previous subsection.

Table 2.4 summarizes the advantages and disadvantages for each technique, described in section 2.3.1. Nevertheless, these advantages and drawbacks may not be similar in all models since each model behaves distinctly depending on its governing equations.

Table 2.4: Advantages and disadvantages of AI models

AI model	Advantages	Disadvantages
ANN	- The most practical AI model. - Ability to deal with minimal input data. - Having a distributed memory. - Can execute multiple tasks simultaneously. - No limitations about the input and output vector relationships.	- There are no standards or criteria for the design and execution of the models. - ANN can only work with numerical information. - No explanation for the test answers.
ANFIS	- Easy to use in both linguistic and numerical knowledge. - It is fast and accurate. - Ability to solve complicated nonlinear problems. - Advantage of using both artificial neural network and fuzzy logic.	- Complexity increases when the amount of fuzzy rules rise. - In the case of absent data, there will be a shortcoming in the system performance.
SVM	- Effective when dealing with high dimensional data. - Can be utilized for both classification and regression problems. - Able to work with image data.	- Not suitable for large data sets as it takes time to train. - SVM is complex as it is difficult to interpret and understand the model. - Not able to explain the classification in terms of probability since SVM is not a probabilistic model

Table 2.4: Continued

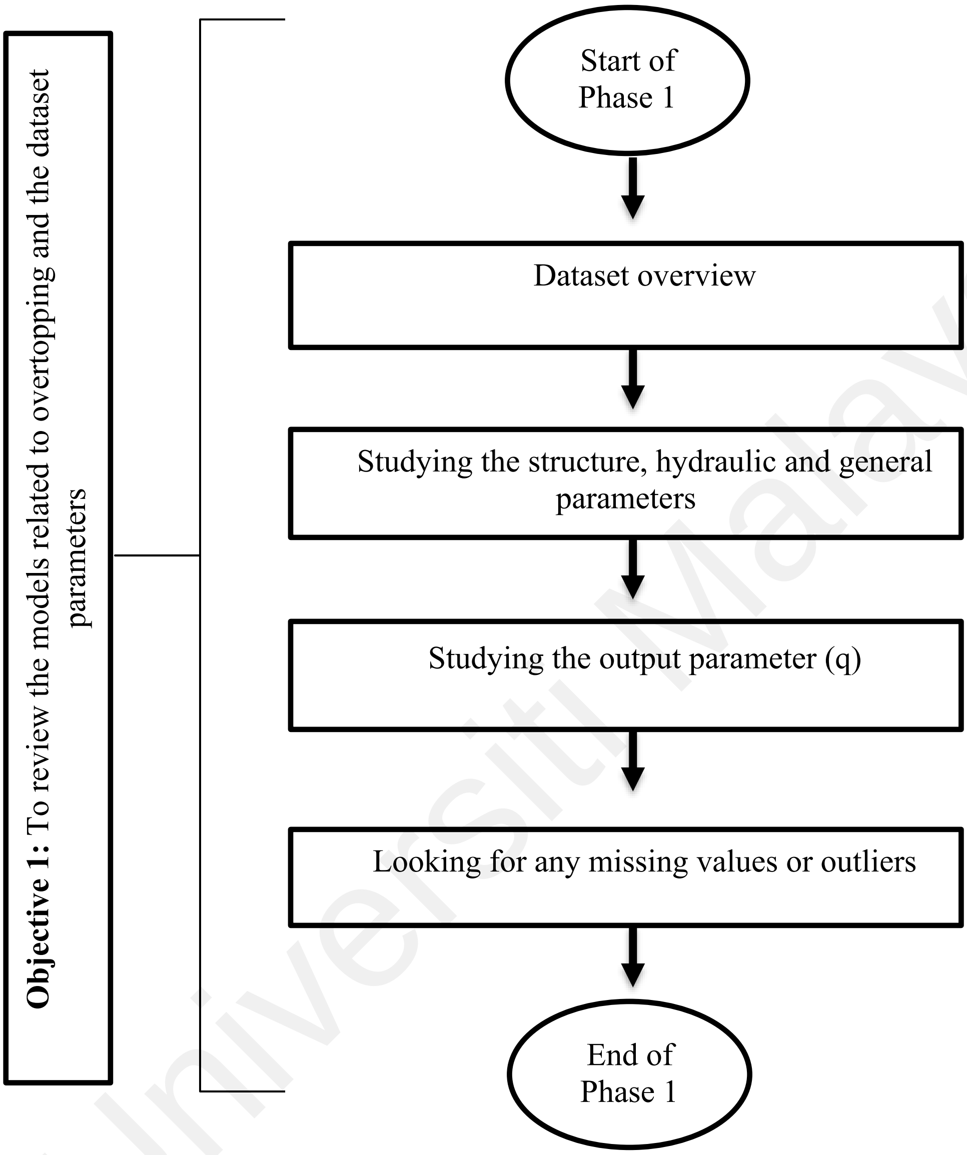
XGB	- Fast and efficient training and prediction times. - Handles missing values and large number of features well. - A wide variety of tuning parameters to control model complexity and improve accuracy. - Robust to outliers and noisy data. - Can be used for both regression and classification problems.	- Can be sensitive to small changes in data. - Can be difficult to interpret compared to other models. - May overfit if not used with proper parameter tuning. - Can be memory intensive for large datasets. - Training time can be longer compared to some other algorithms.

CHAPTER 3: RESEARCH METHODLOGY

3.1 Introduction

Wave overtopping is a significant coastal hazard that can cause damage to infrastructure, erosion of beaches, and flooding. Wave discharge prediction and the dynamics of wave overtopping study are essential for coastal management and protection. In this study, the aim is to develop a novel methodology for predicting wave discharge using artificial intelligence (AI) and conducting a physical experiment to test the new data compared to the original database. The current investigation consisted of two parts, where each part consists of four (4) phases. The first part of the study is to predict wave overtopping through machine learning. To accomplish the research objectives, the study will focus on using the XGBoost model, a machine learning algorithm, to predict wave overtopping discharge. Additionally, the performance of the XGBoost model will be optimized by adjusting its hyperparameters through a process known as model tuning. The second part of the study describes the physical experiment of the study. It involves a physical experiment conducted at NAHRIM using a 2D wave flume to quantify wave overtopping. The collected data then will be added to the existing database. Phase 1 of part 2 aims to develop a robust and reliable model is estimation of wave overtopping. This model will be a crucial tool in assessing potential risks and enhancing safety measures for coastal structures. In phase 2, the insights gained from the developed model will be used to evaluate the impact of incorporating a new dataset. By applying advanced machine learning techniques, we seek to improve the predictive capabilities further and ensure the model adaptability to changing environmental conditions. The results of this study will provide valuable insights into the dynamics of wave overtopping and improve the accuracy of wave discharge predictions, which can add a value to the coastal management and protection strategies. Figure 3.1 shows the flowchart of the research program.

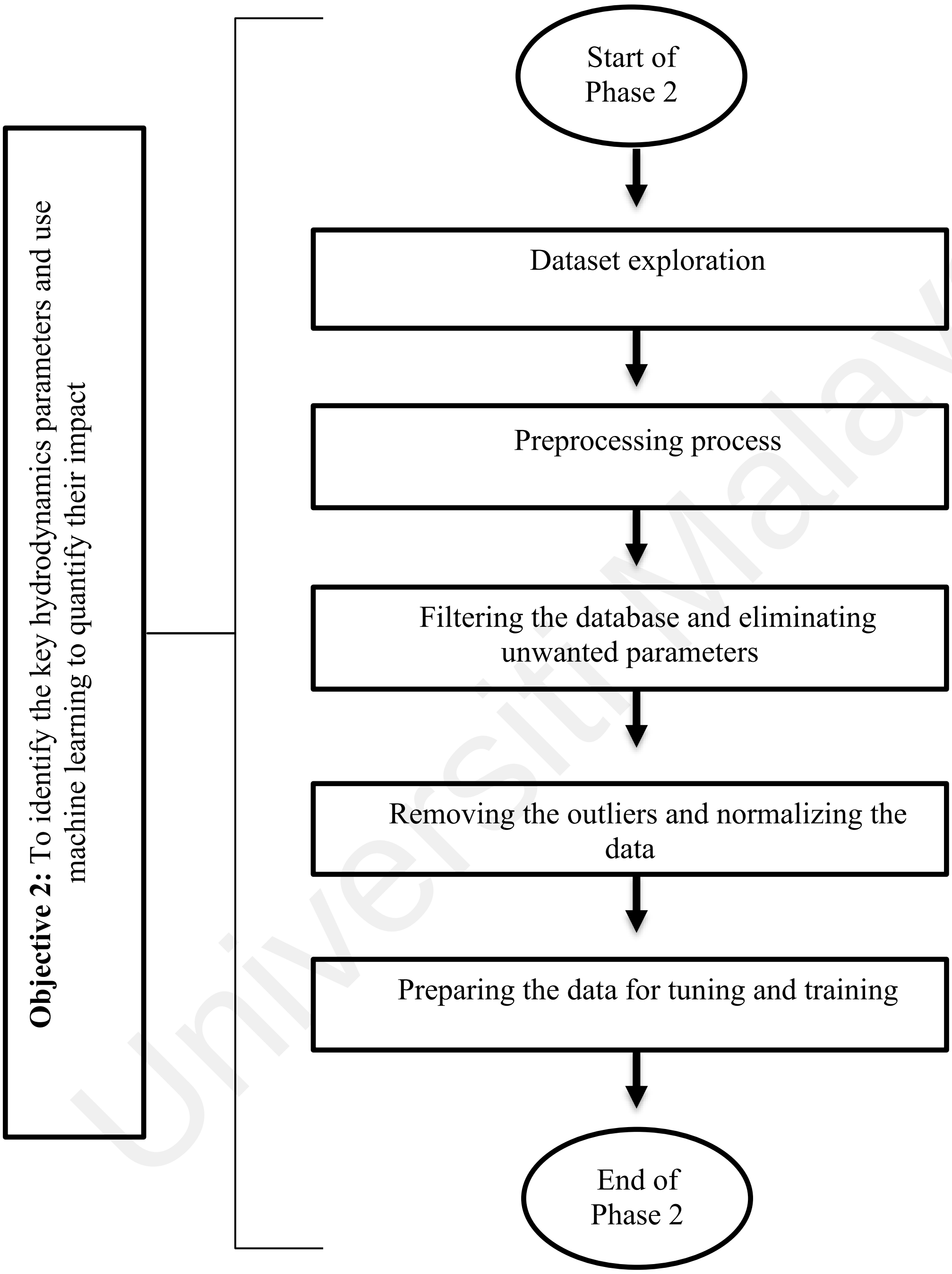
Part 1:


Objective 1: To review the models related to overtopping and the dataset parameters
Start of Phase 1
Dataset overview
Studying the structure, hydraulic and general parameters
Studying the output parameter (q)
Looking for any missing values or outliers
End of Phase 1

(a) Phase 1, part 1

Figure 3.1: Flow chart of the research

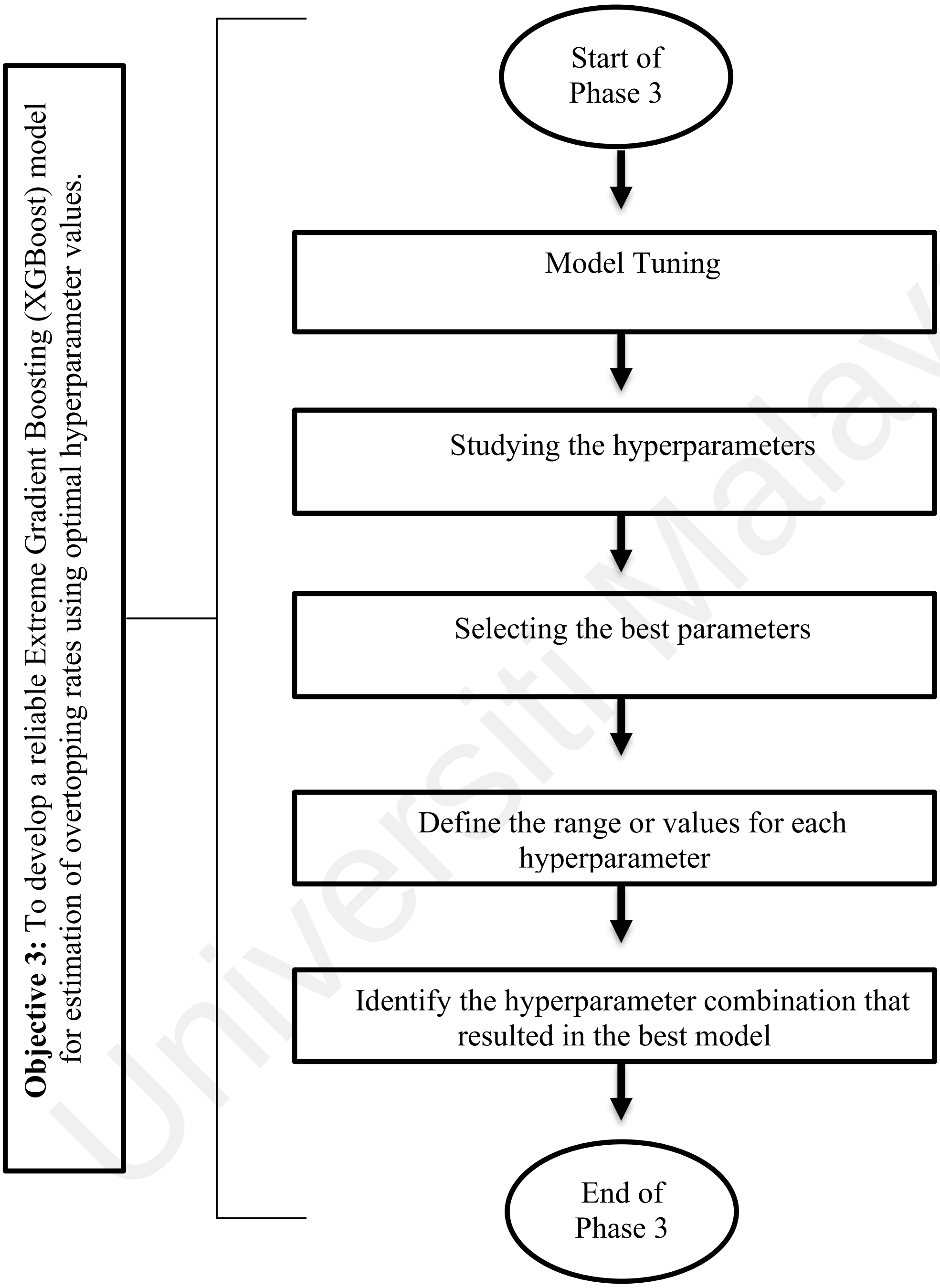
Part 1 cont.:


Objective 2: To identify the key hydrodynamics parameters and use machine learning to quantify their impact
Start of Phase 2
Dataset exploration
Preprocessing process
Filtering the database and eliminating unwanted parameters
Removing the outliers and normalizing the data
Preparing the data for tuning and training
End of Phase 2

(b) Phase 2, part 1

Figure 3.1: Continued

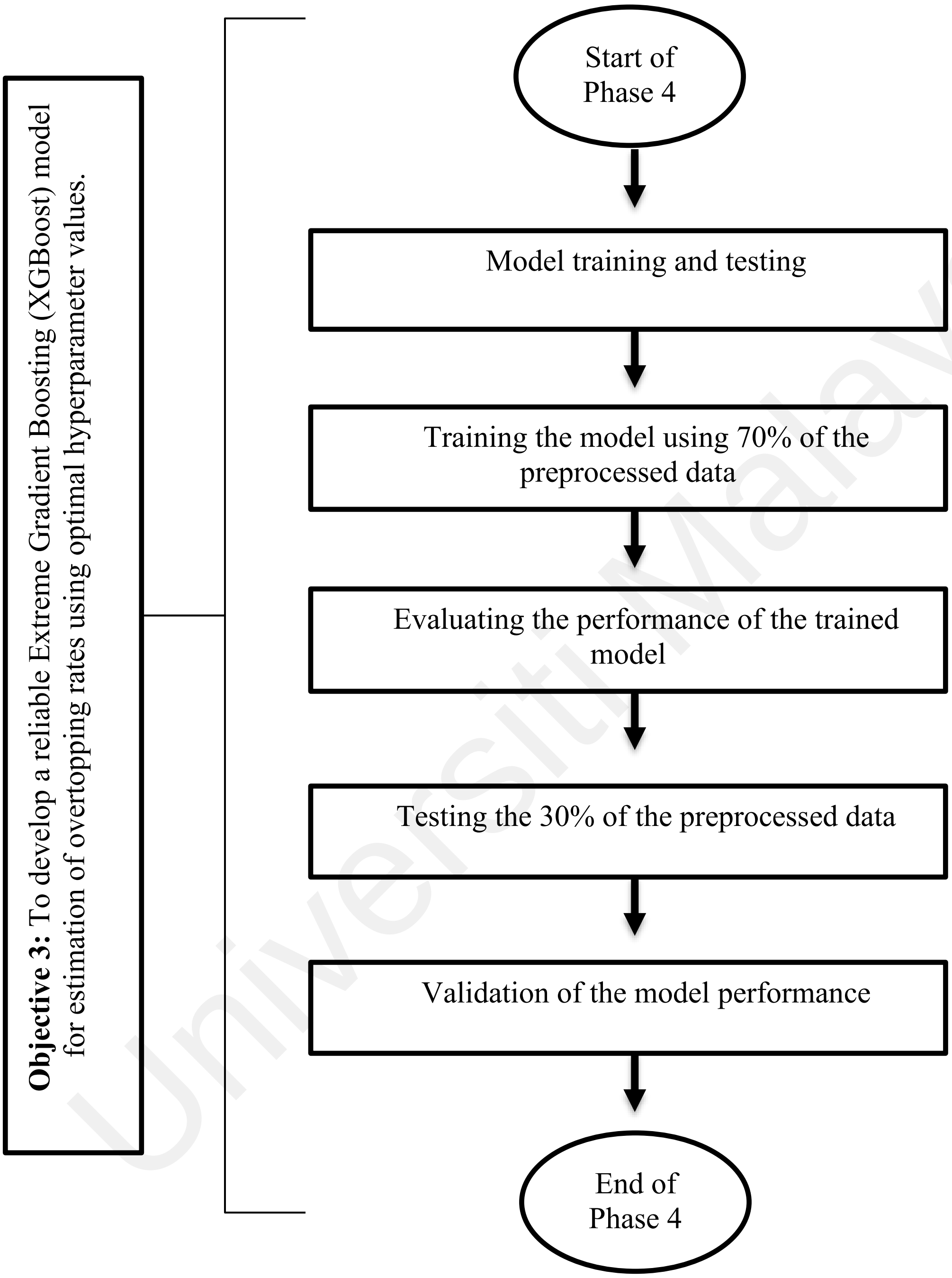
Part 1 cont.:


Objective 3: To develop a reliable Extreme Gradient Boosting (XGBoost) model for estimation of overtopping rates using optimal hyperparameter values.
Start of Phase 3
Model Tuning
Studying the hyperparameters
Selecting the best parameters
Define the range or values for each hyperparameter
Identify the hyperparameter combination that resulted in the best model
End of Phase 3

(c) Phase 3, part 1

Figure 3.1: Continued

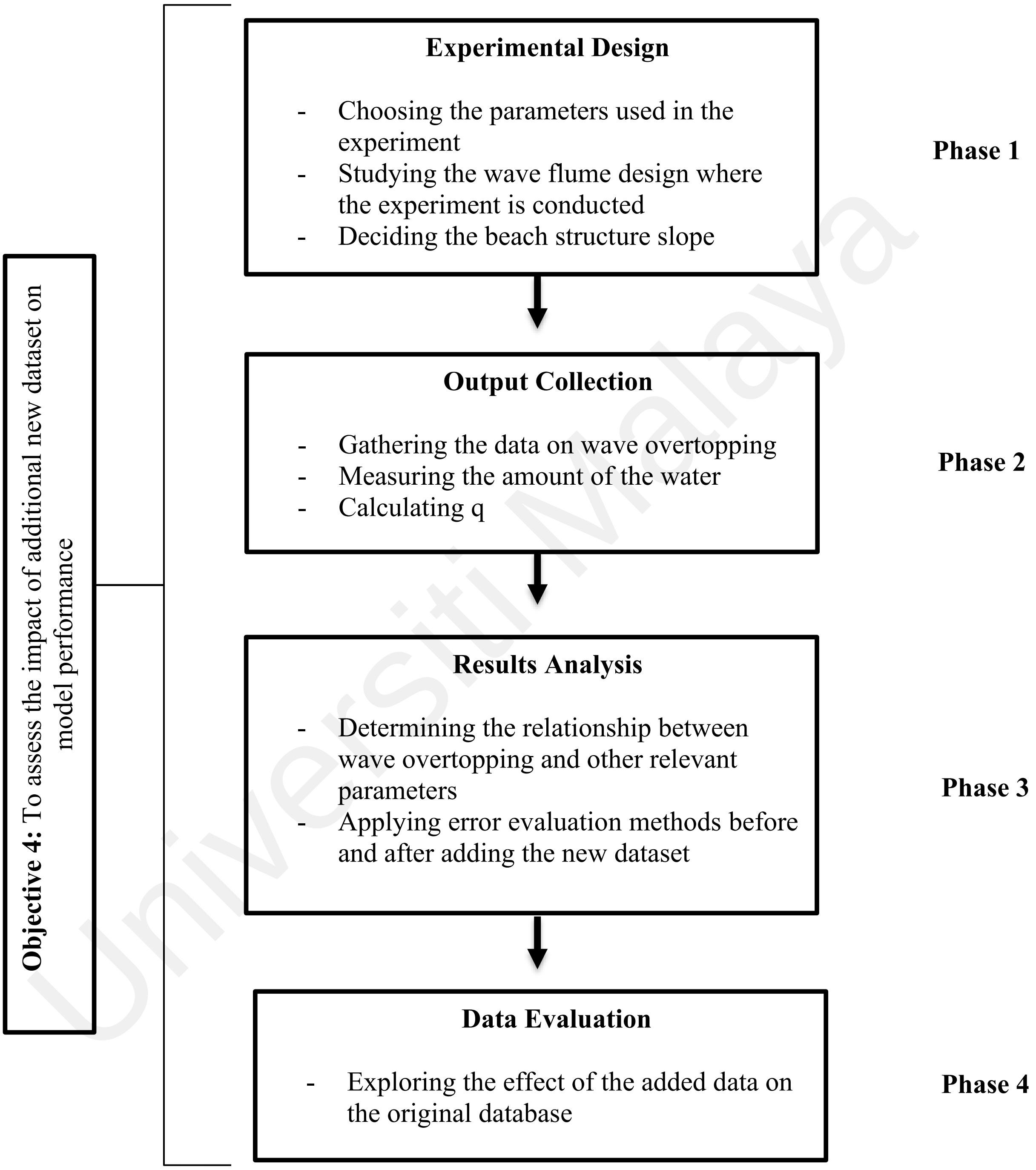
Part 1 cont.:


Objective 3: To develop a reliable Extreme Gradient Boosting (XGBoost) model for estimation of overtopping rates using optimal hyperparameter values.
Start of Phase 4
Model training and testing
Training the model using 70% of the preprocessed data
Evaluating the performance of the trained model
Testing the 30% of the preprocessed data
Validation of the model performance
End of Phase 4

(d) Phase 4, part 1

Figure 3.1: Continued

Part 2:


Objective 4: To assess the impact of additional new dataset on model performance
Experimental Design
- Choosing the parameters used in the experiment
- Studying the wave flume design where the experiment is conducted
- Deciding the beach structure slope
Phase 1
Output Collection
- Gathering the data on wave overtopping
- Measuring the amount of the water
- Calculating q
Phase 2
Results Analysis
- Determining the relationship between wave overtopping and other relevant parameters
- Applying error evaluation methods before and after adding the new dataset
Phase 3
Data Evaluation
- Exploring the effect of the added data on the original database
Phase 4

(e) Part 2

Figure 3.1: Continued

3.2 Development of wave overtopping prediction model

3.2.1 Procedure of gathering overtopping information

To initiate part one of the study, which is prediction model development for wave overtopping, a comprehensive and trustworthy overtopping database was created by collecting all available information from all series of tests. This included not only data on wave characteristics, test structures, and overtopping discharges, but also details on the test facility, measurement processing, and accuracy of the testing process.

Below are some questions related to the overtopping, wave characteristics, test facility, test structure, and processing needed to be answered for every overtopping test:

- What was the specific measurement - the volume of overtopping or the percentage of overtopping waves? and how was the volume of overtopping determined? - through measuring the rise in water level or the weight of the overtopping water?

o What were the wave properties of the storm that was measured or produced, including the type of waves (regular or irregular), the length of the wave crests (long or short), the wave height and period, and the incident wave angle?

❖ What testing facility was utilized? (Wave basin or wave flume and 2D or 3D tests) and what were the capabilities/limitations of the wave generation system? Did reflection compensation occur during testing? Was it active or passive wave absorption? What was the scale of the model used?

- "What was the type of structure tested (e.g. vertical wall, sloping structure, etc.)? What were the geometric specifications of the structure? What materials were utilized in building the test section? What was the appearance of the foreshore?"

➢ "Did the researcher conduct time domain and/or spectral domain analysis? Did they analyze reflections, including separation of incident and reflected waves or simply determining the total waves? How were the incident waves measured - was there a calibration of the testing facility at the structure's location prior to construction, measurement of waves at the structure's toe during testing, or only measurement at deep water?"

Based on the answers to these questions, each test was evaluated for reliability and complexity, which were factored into the database through the creation of a "Reliability Factor" (RF) and a "Complexity Factor" (CF) for each test. The RF reflects the reliability of the test performed, while the CF represents the complexity of the overtopping structure.

3.2.2 Dataset parameters

To utilize the overtopping database for the development of a neural prediction method, each test needed to be characterized by a specific set of parameters. These parameters were carefully chosen to offer a comprehensive representation of the overtopping test. They encompassed the measured overtopping data, the measurement reliability, and the complexity of the structural section. The parameters were categorized into three groups: hydraulic parameters, structural parameters, and general parameters. The hydraulic parameters described the wave characteristics and measured overtopping, while the structural parameters provided information about the test structure. The general parameters included additional relevant details about each overtopping test. For a concise summary of these parameters along with their definition simplifications, refer to Table 1.1 and Figure 1.3 in Chapter 1.1.

More elaborate information about structural parameters, hydraulic parameters and general parameters will be discussed in sections 3.2.2.1, 3.2.2.2 and 3.2.2.3 respectively.

3.2.2.1 Study of structure parameters

Each overtopping structure's first studying stage consists of dividing the structure into three primary sections. The waves that attack the structure are the beginning point here.

The most important part for the waves is the structure surrounding the SWL. This area is either larger or smaller depending on the size of the waves. Referring to van der Meer (1998), the region of the structure between 1.5H_{m0} toe above and 1.5H_{mo} toe below SWL is regarded as the governing part where the wave action is concentrated. The area defined by 1.5 H_{mo} toe above and 1.5 H_{mo} toe below SWL is referred to as the structure's "centre area." The structure's lower part is referred to as the "lower area," while the higher section is referred to as the "upper area."

The top or lower area may be lacking depending on the wave height and water level. Depending on the wave height and water depth near the structure, the centre region can extend the structure slope, but it can also enclose a section of the structure's toe. However, other possibilities still may happen. When looking at the structural parts of coastal structures in general, it is easy to observe:

- a structure body with a vertical wall, a sloping part, or a mixture of the two, and possibly containing a structure berm,
- a structural toe that protects the lower section of the structure structurally, and
- a structure's crest, which serves as a support for the upper part of the structure.

As illustrated in Figure 3.2, a structural berm is most likely located in the structure's centre area (= the area between 1.5Hm0 toe above and 1.5H_{mo} toe below SWL. If the berm is located lower, the waves are more likely to perceive it as a toe. The berm is more likely to be sensed as a crest if it is higher. In relation to the position of the berm, a toe is defined as most likely to appear in the lower section of the structure (= lower than 1.5Hm0 toe below SWL) and a crest as most likely to appear in the upper section of the structure (= higher than 1.5Hm0 toe above SWL). A brief description of the structural parameters is given below.

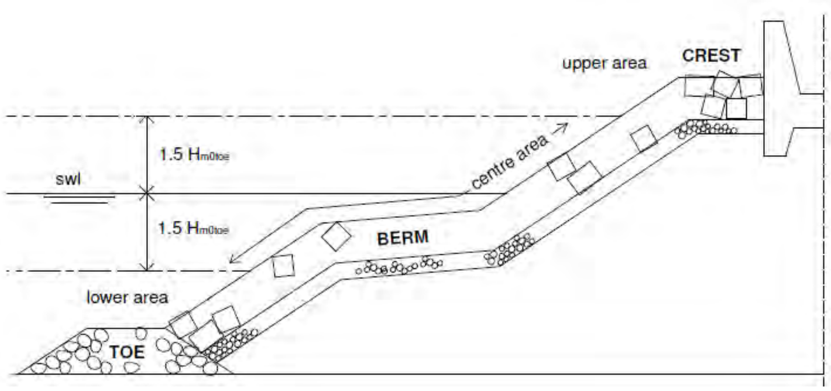


Figure 3.2: Berm, crest and toe usual positions. Retrieved from (Verhaeghe, 2005)

(a) ***h_{deep} [m]***

This is the water depth in deep(er) water. Deep wave characteristics such as $Hm0_{deep}$, Tp_{deep}, Tm_{deep}, and $Tm\ -1,0_{deep}$ are present at this water depth. According to the definition, h_{deep} does not have to be the deepest water depth in the flume or basin for laboratory experiments. The value of h_{deep} is located between the water depth at the toe of the structure and the deepest water depth in the flume, depending on where the wave gauges are located. The deep-water depth corresponds to the water depth in front of the flume's wave paddle.

(b) ***m [-]***

The slope of the foreshore is indicated by the parameter m (measured vertically): m (units measured horizontally). If there is not a uniformly sloping foreshore, the value of m must be estimated. The mean value over a horizontal distance of around 2 wave lengths L_0 in front of the structure is a useful approximation of m. The approximation to the foreshore right in front of the structure can be justified because this part qualifies for the incident wave characteristics. In some cases, the value of m should theoretically be infinite, but because a real, finite value is more practical, a value of 1000 was assigned to m in the database.

(c) ***h [m]***

The water depth right in front of the structure's toe is represented by the value of h. The water depth "at the toe of the structure" is a term used frequently. The value of h equals the value of h_{deep} in the case of a flat flume bottom.

(d) ***h_t [m], B_t [m]***

These are the water depth on the toe and the width of the toe, respectively. The middle of the toe is where h_t is measured. On the top of the toe, the value of B_t is measured. The database does not include the toe's front slope since it appears to be a less important factor considering the toe's overall low water level position. Furthermore, the front slope of a structure toe is frequently 1: 2. As a result, an additional requirement for defining a toe may be that the front slope should approximate 1: 2. If the structure does not have a toe, the water depth on the toe, h_t, is equal to the water depth at the structure's toe, h. in this case, the width of the toe, B_t, is equal to zero.

(e) ***B [m], h_b [m], tanα_B [-], B_h [m]***

These are the four parameters that describe the berm of an overtopping structure. The berm width is measured horizontally and is represented by the value of B. The water depth on the berm is measured in the centre of the berm and is represented by h_b. The value of h_b is negative if the berm is located above SWL. The tangent of the angle formed by a sloping berm with a horizontal is Tanα_B. Tanα_B = 0 if the berm is horizontal.

The breadth of the horizontally schematized berm is represented by B_h. The value of B_h = B when there is a horizontal berm (i.e. $\tan_B = 0$), but in case of a sloping berm, $B_h <$ B. The value of B_h is found by extending the structure's upper and lower slopes up to the level of the berm's centre point. The horizontal schematization of the berm is obtained by connecting these two points. In case the structure has no berm, all the values of B, B_h, tan B, h_b will be equal to zero, except for a composite slope. In the case of a composite slope which is structure made up of successive separate slopes without a horizontal component in between. The transition depth between two successive slopes is defined as $h_{b.}$ Despite the absence of a berm, h_b does not equal zero in this case. Defining h_b as the depth of transition between two consecutive slopes amounts to defining a berm with a berm width and slope of zero at this location.

(f) ***R_c [m], A_c [m]. G_c [m]***

The upper part of an overtopping structure is described by these parameters. R_c is the structure's crest freeboard. It is the vertical distance between SWL and the point on the structure where overtopping is measured. This is not usually the structure's highest point.

The structure's armour crest freeboard is referred to as A_c. It is the vertical distance between SWL and the upper limit of the armour layer in the case of armoured structures. If there are structures without armour, such as vertical structures or smooth slopes, A_c can be combined with R_c and G_c to define the structure's crest in more detail. $A_c = R_c$ is often the case. The crest width is represented by G_c. When there is no wave return wall, the parameter Gc only includes the permeable horizontal component of the crest, as it is assumed that overtopping water simply passes an impermeable surface when it reaches it. If the crest consists of an impermeable horizontal road and overtopping is measured behind a wall on the landside of the road, the crest width Gc will logically be equal to the road's width, as only the water passes the wall will be measured.

(g) ***$\cot\alpha_d$ [-], $\cot\alpha_u$ [-], $\cot\alpha_{excl}$ [-], $\cot\alpha_{incl}$ [-]***

The slope of the overtopping structure is described by these parameters. It is important to note that the structure's toe and crest are not included in these four slope parameters because they are already represented by other parameters. The four parameters can be used to explain the overtopping structure in three different ways:

- with cotα_d and cotα_u or
- with cotα_{excl} or
- with cotα_{incl}

The cotangents of the angle formed by the structure part in the centre area below (cotα_{down}) and above (cotα_{up}) the berm with a horizontal are cot$_d$ and cot$_u$.

Calculated mean slopes are referred to as cotα_{excl} and cotα_{incl}. The cotangent of the structure's mean angle with a horizontal is cotα_{incl}, and the berm (if positioned in the structure's centre region) is included in this mean value (cot$\alpha_{inclusive}$ berm). Cotα_{excl} is the cotangent of the structure's mean angle with a horizontal, without taking into consideration the current berm (cot$\alpha_{exclusive}$ berm). The use of both the cotα_u and cotα_d parameters typically result in a better schematization than using simply one of the mean parameters cotα_{excl} or cotα_{incl}.

The structure's upper slope angle, α_u, relates to the slope above the berm, which is obtained by connecting the structure's point at a level of 1.5$H_{m0\ toe}$ above SWL with the leeside endpoint of the berm. If the structure's crest is located in the centre (less than 1.5$H_{m0\ toe}$ above SWL), the starting point of the crest must then be used to determine α_u instead of the point 1.5$H_{m0\ toe}$ above SWL. The lower slope angle of the structure, α_d, relates to the slope below the berm, which is obtained by connecting the structure's point at a level of 1.5$H_{m0\ toe}$ below SWL with the berm's coastal endpoint. If the structure's toe is in the centre (which means it's less than 1.5$H_{m0\ toe}$ below SWL), the starting point of the toe must be used instead of the point at 1.5$H_{m0\ toe}$ below SWL to determine α_d.

The mean slope angle, α_{incl}, is found by connecting the point on the upper slope $1.5H_{m0\ toe}$ above SWL with the point on the lower slope $1.5H_{m0\ toe}$ below SWL. The subscript 'incl' indicates that if there is a berm, it is included in the $\cot\alpha_{incl}$ value.

The mean slope angle, α_{excl}, is determined by deducting the horizontal width of the berm, B_h, from the horizontal distance between the two points that determine α_{incl}, and then dividing the result by the vertical distance between the two points that determine α_{incl}.

(h) ***Y_f [-]***

The permeability and roughness of the structure are indicated by the parameter Y_f. The lower the overtopping, the rougher and more permeable the structure is, due to the loss in wave energy. This is reflected in the lower value of the Y_f parameter.

The introduction of a roughness reduction factor for various types of revetments originates from Russian studies conducted in the 1950s with regular waves and is based on a value obtained for wave run-up. (TAW, 2002) presents more recent values for Y_f for numerous revetment types, based on additional run-up tests with irregular waves, also performed on a wide scale, from 1974 to 2002. In the case of an impermeable smooth construction, (TAW, 2002) proposes a value of 1 for Y_f and a value of 0.7 or 0.55 in the case of one or two layers of rock on an impermeable core.

3.2.2.2 Study of hydraulic parameters

12 hydraulic parameters, which were listed in Table 1.1, are used to define wave characteristics and measured overtopping. Several of these parameters were frequently unavailable from the test report because they were not measured or written down while the test was being performed. The following situations can be recognized in this context:

- Only wave characteristics from deep water were accessible; wave characteristics from the structure's toe were absent.
- Deep water wave characteristics were unavailable, and only wave characteristics near the structure's toe were available.
- To identify the wave characteristics, only time domain analysis was performed.
- At either deep or shallow water, only one or two of the three spectral wave periods were available.
- The percentage of waves that overtook P_{ow} was not measured.

For the estimating of characteristic wave parameters in relatively deep water, Longuet-Higgins, (1952) showed that the wave heights of these waves obey the Rayleigh distribution. According to this distribution function, the probability that an individual wave height H exceeds some arbitrary value referred to as H_d (with d < design), in the storm characterized by the root-mean-square wave height H_{rms}, can be expressed by:

$$\boldsymbol{P\ (H > H_d)_{H_{rms}} = \exp\left[-\left(\frac{H_d}{H_{rms}}\right)^2\right]} \qquad \textbf{(3.1)}$$

One can also state that in case of deep water waves with a narrow energy spectrum, all characteristic wave heights are theoretically proportional to the standard deviation of the surface elevation with known proportionality constants. Starting from $H_{rms} = \sqrt{8m_0}$, one also has $H_{1/3} = 4\sqrt{m_0}$ etc. When estimated by m_0 (spectral domain analysis), the notation H_{m0} should be used for the significant wave height:

$$\boldsymbol{H_{m0} = 4\sqrt{m_0}} \qquad \textbf{(3.2)}$$

where m0 is a measure of the total energy of the storm.

While, in shallow water, the wave heights no longer obey the Rayleigh distribution. Shoaling, triad interactions and depth-induced breaking become relevant, causing a profile distortion to the linear deep water waves. The consequence is that the approximation $H_{m0} = H_{1/3}$ is no longer valid in shallow water.

Contrary to the wave height, the wave period of deep water waves does not exhibit a universal distribution law such as the Rayleigh distribution. Nevertheless, it has been empirically found that characteristic period parameters are interrelated at deep water.

With the goal of creating a reliable database, if possible, an acceptable value was searched for these missing parameters or some of the assumptions from previous research and extra calculations were used to achieve this. Nevertheless, in some cases it was simply not possible to estimate missing hydraulic parameters accurately, thus preference was given to leave the value of the missing parameter blank in the database. However, the preprocessing methods applied in this study to prevent such cases in the dataset are mentioned and described in section 3.2.3.

3.2.2.3 Study of general Parameters

For each overtopping test, the database contains two general parameters: CF and RF. This section discusses how the values for these parameters are assigned.

(a) The complexity factor CF

The variable of complexity CF is the factor that indicated the complexity of the overtopping structure. The factor denotes the degree of approximation gained by description of a test structure using database structural parameters. Table 3.1 shows the range of values for the complexity factor CF with a brief explanation provided for each value.

Table 3.1: Complexity factor (CF) range of values

CF	Meaning
1	Simple section: The structural parameters accurately (or as accurately as possible) define the section.
2	Quite simple section: The structural parameters accurately define the section, though not perfectly.
3	Quite complicated section: The structural parameters accurately characterize the section, yet there are certain challenges and uncertainties.
4	Very complicated section: The section is too complex to be described using structural parameters, and the section's description using them is inaccurate.

(b) The reliability factor RF

The reliability factor RF shows the reliability of the considered overtopping test. Table 3.2 explains the range of values for the reliability factor RF. A brief explanation is provided for each value. Also, there are several elements which influence the RF reliability factor:

- The accuracy of the researcher measurements and analysis during the overtopping test
- The restrictions of the test facility utilized to conduct the test
- The estimations/calculations that had to be made due to lacking values

Table 3.2: Reliability factor (RF) range of values

RF	Meaning
1	extremely reliable test: All necessary data is available, and measurements and analysis were carried out in a satisfactory way.
2	Reliable test: Although certain estimates/calculations had to be made and/or some measurements/analysis uncertainties exist, the overall test can be characterized as ‘credible.'
3	Less reliable test: Some assumptions had to be made, and there were some doubts in measurements/analysis, resulting in the test being classified as ‘less trustworthy.'
4	Unreliable test: There were no acceptable estimations/calculations, and/or measurements/analysis contained flaws, resulting in an unreliable test.

3.2.3 Data exploration

Data exploration is a crucial step in the machine learning process, where one analyzes and summarizes the main characteristics of the data. During this process, important parameters such as count of data points, mean value, standard deviation, maximum value, and minimum value are calculated. These parameters give a quick overview of the distribution of the data and help identify potential outliers or anomalies. The mean value represents the average of the data, while the standard deviation measures the spread of the data around the mean. The maximum and minimum values represent the highest and lowest values in the data, respectively. These parameters are used to normalize the data and improve the accuracy of the machine learning models. Table 3.3 describes all the data exploration for the input parameters.

Table 3.3: Data Exploration

Parameters	Hm0 d	Tp d	Tm d	Tm1,0 d	hdeep	h	Hm0 toe
count	5670	5670	5670	5670	5670	5670	5670
mean	0.195	1.9089	1.5888	1.7487	2.2103	0.5564	0.16865
std	0.429	1.2573	1.0455	1.1379	11.822	0.5816	0.27809
min	0.026	0.7272	0.5923	0.6610	0.1203	0.029	0.02071
25%	0.1	1.384	1.1565	1.2616	0.545	0.3	0.09289
50%	0.129	1.652	1.377	1.5150	0.7	0.47	0.1223
75%	0.16	1.9745	1.6300	1.8	0.8	0.61	0.15
max	5.51	15	12.5	13.636	100	5.01	2.403
Parameters	**Tm toe**	**Tm1,0t**	**ht**	**Bt**	**cotad**	**cotau**	**cotaexcl**
count	5670	5670	5670	5670	5670	5670	5670
mean	1.582	1.7926	0.5349	0.0379	2.1213	3.7566	2.06852
std	1.014	1.1599	0.5830	0.1009	1.3359	10.042	1.49580
min	0.609	0.6653	0.029	0	0	-5	-1.3313
25%	1.156	1.2616	0.3	0	1.3333	1.25	1.25
50%	1.377	1.521	0.44	0	2	2	2
75%	1.63	1.9035	0.61	0	3	3	2.88964
max	10.79	10.64	5.01	1	7	100	7.86324
Parameters	**D50_d**	**gf_u**	**D50_u**	**gf**	**D**	**Rc**	**B**
count	5670	5670	5670	5670	5670	5670	5670
mean	0.036	0.7496	0.0364	0.7453	0.0360	0.2531	0.13170
std	0.148	0.2855	0.1486	0.2748	0.1486	0.5256	0.28221
min	0	0.38	0	0.38	0	0	0
25%	0	0.4	0	0.4	0	0.1011	0
50%	0	1	0	1	0	0.15	0
75%	0.042	1	0.042	1	0.04	0.2105	0.15
max	1.25	1	1.25	1	1.25	4.4928	2

Table 3.3: Continued

Parameters	tanaB	Bh	Ac	Gc	β	mmm	RF
count	5670	5670	5670	5670	5670	5670	5670
mean	0.001	0.1229	0.2404	0.1597	2.719	457.85	1.4822
std	0.0099	0.2524	0.5279	0.5753	9.609	465.92	0.4997
min	0	0	-0.03	0	0	7.6	1
25%	0	0	0.1	0	0	50	1
50%	0	0	0.132	0	0	111.11	1
75%	0	0.15	0.2	0.1777	0	1000	2
max	0.0896	2	4.4928	4.8	80	1000	2
Parameters	**WF**	**gf_d**	**Spread s**	**Tp toe**	**cotaincl**	**hb**	**CF**
count	5670	5670	5670	5670	5670	5670	5670
mean	7.2914	0.732	0.289	1.9233	2.4102	0.0175	1.1020
std	1.62819	0.283	1.286	1.2735	1.6644	0.1155	0.302
min	4	0.38	0	0.7318	-1.331	-0.266	1
25%	6	0.4	0	1.384	1.5	0	1
50%	6	1	0	1.652	2	0	1
75%	9	1	0	1.9941	3.2836	0	1
max	9	1	10	13.7	11.299	1.09	2

3.2.4 Data preprocessing

Data preprocessing is an important step in the data analysis process as it helps to ensure the quality and accuracy of the data being used. This step involves the transformation of the data into a format suitable for analysis. The goal of preprocessing is to make the data suitable for use in machine learning algorithms or statistical models. Additionally, preprocessing can also improve the interpretability of the results and help in identifying patterns and trends in the data. Overall, proper data preprocessing is crucial for obtaining meaningful and accurate results in data analysis.

The Eurotop database consisted of approximately 17,000 tests, with only 13,500 for wave overtopping only, each of which was intended to be used to develop a machine learning model. However, after careful evaluation, it was determined that some of these tests were not reliable and could not be used in the model. In order to ensure that the results of the study were accurate and trustworthy, it was necessary to perform a preprocessing step to remove these unreliable tests from the database. Following the previous work by van Gent et al., (2007), the parameter Weight Factor (WF) is utilized which is based on the RF and CF values. The formula used for the weight factor is WF= (4 – RF) . (4 – CF). In this study, the weight factor (WF) is a crucial aspect in determining the likelihood of a test being selected. It ranges from 0 to 9, depending on the values of the reliability factor (RF) and the complexity factor (CF), which can vary between 1 and 4. When WF is higher, there is a better chance of selecting a test. This is because tests are considered reliable and straightforward. Hence, when RF and CF values increase, WF also goes up, indicating the test is more likely to be chosen. RF and CF serve as measures of the reliability and complexity of a test, respectively, and are used to determine the WF. The reliability factor evaluates the consistency and accuracy of the results obtained from the test, while the complexity factor assesses the complexity of the structure. Tests that are reliable and straightforward are deemed to be more valuable in terms of the information they provide. As a result, they are given more weight and are more likely to be selected over tests that are deemed unreliable or complex. The WF provides a way to evaluate the overall worth of a test, making it an essential aspect of the testing process. The range used in this study for WF and for output variable (q) is summarized in Table 3.4. After the preprocessing step was completed, the finalized number of dataset remaining was 5,670. This dataset was used for both training and testing purposes for the development of the model. This was a critical step in ensuring the validity and robustness of the machine learning model that was being developed. The use of a large and reliable dataset is essential for the development of accurate machine learning models, and the preprocessing step was crucial in achieving this goal.

Table 3.4: Range of parameters

Parameter	Range
WF	6-9
q	>10^{-6}

3.2.5 Model tuning and scaling

In the field of machine learning especially XGB algorithm, model tuning play a crucial role in developing accurate and high-performing predictive models. The success of a machine learning algorithm heavily relies on finding the optimal set of hyperparameters and preprocessing techniques to enhance the model predictive capabilities. This section focuses on the tuning process for the XGBoost model, employing the 'scikit lean' library for model evaluation. The first subsection discusses hyperparameter tuning, which involves the systematic search for the best configuration of model parameters. The second subsection explores the application of feature scaling techniques for the input and output to ensure the fairness and effectiveness of the XGBoost model parameter values.

3.2.5.1 Hyperparameter tuning

Hyperparameter tuning in XGBoost involves the systematic process of finding the optimal values for various hyperparameters that control the behavior and performance of the XGBoost model. By tuning these hyperparameters, the study aims to improve the model accuracy and prevent overfitting issue. The process of tuning was conducted using High Performance Computing (HPC) machines at the Heinrich Heine University Düsseldorf, Germany.

The tuning process involves techniques such as grid search, random search, or more advanced optimization algorithms like Bayesian optimization. These methods explore different combinations of hyperparameter values and evaluate their impact on the model performance using a validation dataset. Hyperparameter tuning in XGBoost involves adjusting several key hyperparameters to optimize the model performance. After thorough experimentation with various hyperparameters, the selection of these specific parameters is based on their ability to yield the best results for the model. In this study, a multiple of hyperparameter values are used during the hyperparameter tuning process for several reasons. First, it enables the exploration of various optima, and increases the likelihood of finding optimal or near-optimal values. Second, the model becomes more robust, adapting well to different data distributions and problem contexts. Third, this approach helps identify outliers or unconventional yet effective hyperparameter values that may significantly impact the model performance. Additionally, the perception of model behavior improves and guides future development due to the study of multiple hyperparameters. Careful evaluation and comparison of different hyperparameter combinations led to the identification of the optimal values, which consistently demonstrated superior performance in terms of accuracy and other relevant metrics.

The chosen parameters with the wide range of values, in Table 3.5, reflect an informed decision aimed at optimizing the model performance and achieving optimal outcomes. These hyperparameters are the maximum tree depth (max_depth), minimum number of data points in a leaf (min_child_weight), learning rate (learning_rate), L2 regularization parameter (lambda), and subsampled fraction of training data (subsample) which play a crucial role in shaping the behavior and accuracy of the XGBoost model. The max_depth controls the depth of each tree and helps prevent overfitting, with a default value of 6 and a typical range of 3 to 10. The min_child_weight sets the minimum number of data points required in each leaf node, regulating model complexity, and has a default value of 1, typically ranging from 1 to 20. The learning rate determines the step size at each boosting iteration, influencing the trade-off between convergence speed and generalization. Its default value is 0.3, with typical values ranging from 0.01 to 0.2. The L2 regularization parameter, lambda, adds a penalty term to the loss function, reduces model complexity and overfitting, and has a default value of 1, with a range of 0 to 10. The subsample hyperparameter controls the fraction of training data used for each boosting iteration, to introduce randomness and enhance generalization. Its default value is 1 (indicating no subsampling), while typical values range from 0.5 to 1. The exploration of different values within these ranges and the evaluation of their impact are crucial to find the optimal combination that maximizes the XGBoost model performance for a specific problem and dataset. In Table 3.5, the optimal value for each hyperparameter, which shown in red, indicates the settings that led to the best performance and generalization of the machine learning model. For example, the value for max depth indicates that the ideal depth or maximum number of levels in the tree that should be grown during the training process is 21. Also, using the value 0.75 for subsample ensures that the model is trained with sufficient randomness to generalize well with no overfitting on the training data.

Table 3.5: The hyperparameters with the values utilized.

Hyperparameter	Name	Values
Learning rate	max_depth	3; 4; 5; 6; 7; 10; 15; 17; 20; (21); 22
L2 regularization parameter	min_child_weight	1; (3); 5; 7; 10; 12; 15; 16; 18; 19; 20
Subsampled fraction of training data	learning_rate	0.0005; 0.0001; 0.005; 0.0075; 0.01; 0.05; (0.07)
Maximum tree depth	reg_lambda	(1); 2; 3; 4; 5; 7
Minimum number of data points in a leaf	subsample	0.15; 0.25; 0.5; (0.75); 0.85

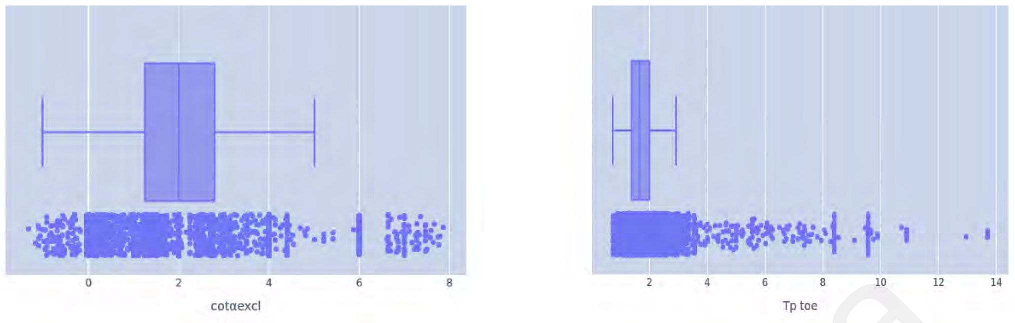
3.2.5.2 Standardization and scaling

Standardization process, also known as feature scaling, is implemented on all the input parameters as a part of improving the accuracy model. It is a crucial data preprocessing technique employed in machine learning and data analysis to ensure that features or variables exhibit comparable scales and distributions. Standardization is a scaling technique that centres the values of a variable around the mean and adjusts them to have a standard deviation of one. The process involves the following steps. First, the mean (μ) and standard deviation (σ) of the feature are calculated. This provides information about the average value and the spread of the data points. Second, each data point is subtracted by the mean, which centres the data around zero. This step ensures that the new mean of the transformed data becomes zero even if the actual value of the parameter is not zero, i.e., wave period, wave height and etc. This occurs because standardization involves subtracting the mean of the feature from each value and then dividing the result by the standard deviation. Therefore, when the original values are greater than the mean, the standardized values will be positive, while if the original values are lower than the mean, the standardized values will be negative. If a particular feature has a mean greater than zero, after standardization, the mean of that feature will be transformed to zero. Consequently, the standardized values that were originally above the mean will become positive, and those that were below the mean will become negative. However, it is important to note that these standardized values do not represent the actual values of the feature, but rather the number of standard deviations each value deviates from the mean. Thus, even if the original values of a feature cannot be zero, the transformation of standardization can result in the standardized values of that feature being cantered around zero. By standardizing the data, the distribution of values is transformed to follow a standard normal distribution, with a mean of zero and a standard deviation of one. Such process is primarily intended to achieve comparable scales and facilitate fair comparisons between features, rather than representing the original values themselves. Below is the formula used for standardization:

$$\boldsymbol{z} = \frac{\boldsymbol{X} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{3.3}$$

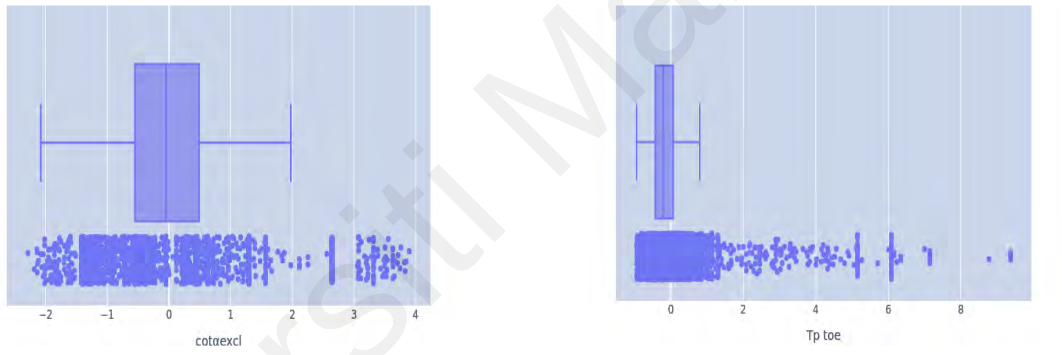
Where z represents the standardized value, x represents data value, μ represents the mean and σ represents the standard deviation.

The process of standardization involves several key steps. Initially, the dataset comprising the target variables of interest is collected. Subsequently, relevant libraries are imported, in this study, the widely utilized scikit-learn library (commonly referred to as 'sklearn'). The scikit-learn library offers an extensive suite of machine learning tools, including preprocessing functionalities like standardization. Upon library importation, the dataset is partitioned into a feature matrix (X) and, when applicable, a target vector (y). The feature matrix (X) represents the independent variables or input features, while the target vector (y) signifies the dependent variable or output. The ensuing step involves the application of standardization to the feature matrix (X) utilizing the 'StandardScaler' class from the 'sklearn.preprocessing' module. This class encompasses the 'fit_transform()' method, which performs two primary operations: the calculation of the mean and standard deviation for each feature and subsequently data adjustment based on these statistical measures. When the 'fit_transform ()' is applied to the feature matrix (X), the resulting output ('X_scaled') represents a standardized rendition of the data. Subsequent to the standardization phase, the standardized feature matrix ('X_scaled') is utilized to train the machine learning model, in conjunction with the target vector (y) if relevant. It is critical to emphasize that the standardized data should be employed for all aspects of model training, evaluation, and prediction. Standardization accomplished through the 'sklearn' library ensures that each feature possesses a mean of zero and a standard deviation of one. This transformation ensures comparability among variables, mitigates biases arising from dissimilar scales, and facilitates the convergence of models utilizing certain algorithms or distance-based computations.



(a) (b)

Figure 3.3: Tp, toe and cotαexcl before applying the standardization



(a) (b)

Figure 3.4: $T_{p,\ toe}$ and cotα_{excl} after applying the standardization.

Figure 3.3 and Figure 3.4 show the difference in scaling for some of the parameters before and after applying the scaling technique These figures illustrate the impact of standard scaling on peak wave period (T_p) and cotangent of the mean angle excluding the berm (cotα_{excl}). The range of values is compressed around zero value, aligning with the goal of standardization. It became apparent that the values varied significantly in terms of its ranges (Figure 3.3). Recognizing the importance of mitigating any bias resulting from these disparate scales, a decision was made to standardize the scale across all features. This standardization process aimed to ensure fairness in the modeling by placing all features on a comparable scale, regardless of their original ranges (Figure 3.4). By standardizing the scales, the model could effectively evaluate the relative importance and contribution of each feature without being influenced by their differing magnitudes. It is important to note that in the presence of outliers, some values may still fall outside the desired range.

The wave overtopping data might exhibit a wide range of values, with some instances having much higher magnitudes than others. In such cases, the data distribution may be skewed, and the extreme values could disproportionately influence the model performance (Figure 3.5). By applying a logarithmic transformation to the output feature, the range of values was compressed, to emphasize the differences in the lower range and lower the impact of outliers in the upper range, as illustrated in Figure 3.6. This process will lead to a more balanced and true representation of the data and enhance the XGB model capabilities. By incorporating logarithmic scaling for the q while standardizing the input features, a consistent framework for the XGB model can be created.

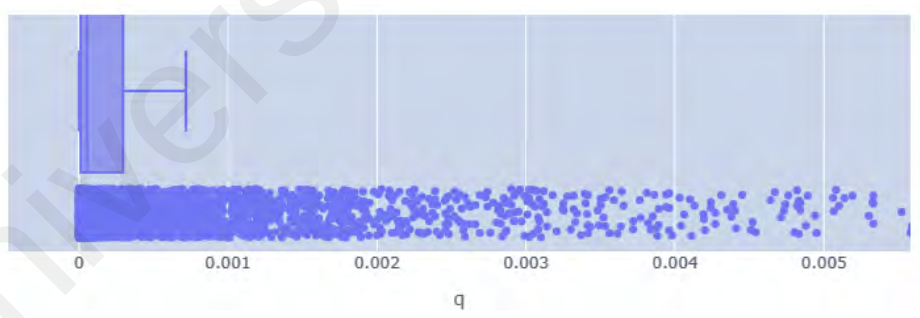


Figure 3.5: Distribution of q feature before log q scaling

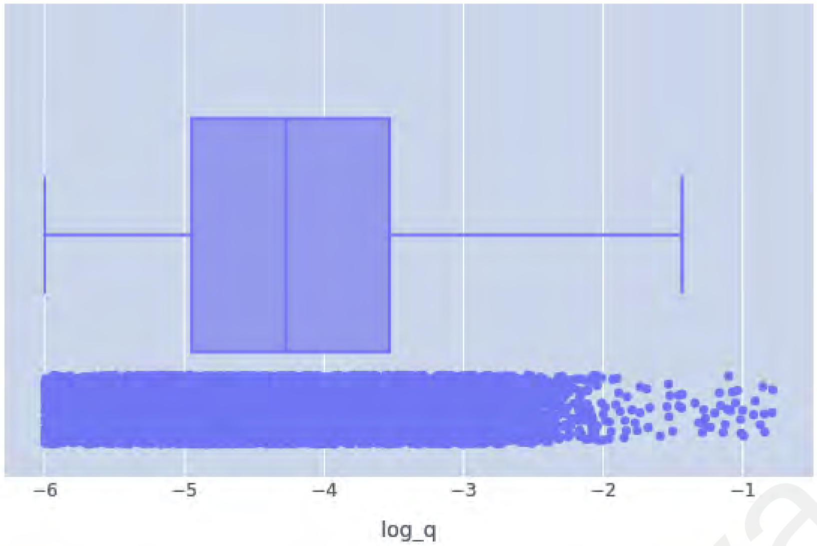


Figure 3.6: Distribution of q feature after log q scaling

Initially, the model sensitivity to small values was inadequate, hence impacting its ability to effectively detect such values. To address this issue, the log scale was applied to the q values. The log scale, a mathematical method, transforms small values into negative numbers on a larger scale, thereby amplifying their magnitudes. Consequently, after the application of the log scale, the values of 'q' ranged from -1 to -6, as evident from the transformed values. Figure 3.5 demonstrates that the differences between the original values of 'q' were minimal due to their small magnitudes. This limited differentiation made it challenging for the model to effectively capture variations and make accurate predictions. However, after the application of the log scale, as depicted in Figure 3.6, the differences between the transformed values became more pronounced. This transformation expanded the range of the values, making them more distinguishable to the model. Consequently, the model could achieve better predictions by using the enhanced information from the log-scaled 'q' values.

The standardized input features ensure that each feature contributes equally to the learning process, while the log scaling of the q mitigates the influence of extreme values and enhances the model ability to capture patterns within a wider range of magnitudes.

This combined approach can potentially lead to better performance, and reliable predictions of wave overtopping rate.

3.2.6 Model training

Model training is a crucial phase in the machine learning pipeline that follows the preprocessing phase, hyperparameter tuning and scaling. This section delves into the training process of the XGBoost model, utilizing the popular 'xgboost' library for model training. During model training, the XGB algorithm learns from the preprocessed data to make accurate predictions on new, unseen data. The preprocessed dataset is typically divided into training and testing sets to enable the model to learn from the training data and then assess its performance on the validation data. During model training, the model is repeatedly adjusted using scaled training data. This adjustment is to minimize the difference between the predicted and actual values (i.e., the loss function). This optimization process is carried out for a predefined number of epochs or until the model converges. Properly tuned hyperparameters from section 3.2.5.1 significantly impact the training process, as they dictate the learning rate, regularization, and other aspects that affect the model convergence and performance. For the training process, 70% of the total tests, amounting to 3970 tests, were utilized. Also, 36 carefully selected parameters were chosen after the filtration process to be employed for both training and testing. The details of these parameters can be found in Table 3.6. After successful model training, the resulting model can be evaluated on a separate test dataset to measure its performance on unseen data and validate its effectiveness in solving the target problem.

Table 3.6: Summary of the features used in the XGB model and its definition

#	Parameter	Unit	Definition of the parameter	Type
1	$H_{m0\ deep}$	m	Off-shore significant wave-height	hydraulic
2	$T_{m\ deep}$	s	Off-shore average wave period	hydraulic
3	$T_{m-1,\ deep}$	s	Off-shore spectral wave period	hydraulic
4	$T_{p\ deep}$	s	Off-shore peak wave period	hydraulic
5	H_{deep}	m	Water depth at the structure toe	structural
6	A_c	m	Wall height with respect to SWL	structural
7	D	m	Average size of the structure elements in the run-up/down area	structure
8	$H_{m0\ t}$	m	Significant wave-height at the structure toe	hydraulic
9	$T_{m-1,\ t}$	s	Spectral wave period at the structure toe	hydraulic
10	$cota_u$	-	Cotangent of the angle that the part of the structure below/above the berm makes with a horizontal	structural
11	$cota_d$	-		structural
12	B_t	m	Toe width	structural
13	h_t	m	water depth on the toe of a structure	structural
14	$cota_{incl}$	-	Cotangent of the mean angle that the structure makes with a horizontal, including/excluding the berm, in the run-up/run-down zone	structural
15	$cota_{excl}$	-		structural
16	γ_f	-	Roughness factor [average in the run-up/down area in the new database	structural
17	S	-	Spreading factor	structural
18	G_c	m	Crest width	structural
19	h_b	m	Berm submergence	structural
20	B_h	m	Horizontal berm width	structural
21	B	m	Berm width	structural

Table 3.6: Continued

23	mmm	-	Foreshore slope, 1: m	structural
24	β	-	Angle of wave attack	hydraulic
25	h	m	Water depth at the structure toe	structural
26	$T_{p,toe}$	s	Peak wave period at the structure toe	hydraulic
27	Tm, toe	s	Average wave period at the structure toe	hydraulic
28	γ_{fd}	-	Roughness factor for cotαd	structural
29	γ_{fu}	-	Roughness factor for cotαu	structural
30	D_d	-	Size of the structure elements along cotαd	structural
31	D_u	-	Size of the structure elements along cotαu	structural
32	R_c	m	Crest height with respect to SWL	structural
33	CF	-	'Complexity Factor' The complexity of the test is indicated by a score with a possible range of 1 to 4.	general
34	RF	-	'Reliability Factor' The reliability of the test is indicated by a score with a possible range of 1 to 4.	general
35	WF	-	Weight factor	general
36	q	m^3/s/m	Average specific wave overtopping discharge	hydraulic

3.3 Physical experiment

3.3.1 Wave flume

For part two of the study, the physical experiment was conducted in a 2D wave flume at National Water Research Institute of Malaysia (NAHRIM) to determine the wave overtopping. The wave flume used for this experiment has dimensions of 50m in length, 1.5m in width, and 2m in height (Figure 3.7). This wave flume is equipped with a wave generator that can generate both regular and irregular waves, which can be adjusted to simulate a range of different wave conditions. Additionally, the wave flume features active wave absorption, to dissipate the wave energy and prevent reflections that could interfere with the experiment. The wave energy spectrum used in this experiment is based on significant wave height (Hs) and peak wave period (Tp). The wave generator has a maximum wave generating capability of 0.5m and can generate waves with periods ranging from 1 to 5 seconds at a maximum flow of 1000 l/s. The control signal for the wave generator is obtained from computer software, which provides a random signal with a predetermined wave energy density spectrum. This allows the experiment to simulate a range of different wave conditions that are representative of the real world. The wave flume can also be used to study a range of other important aspects of coastal engineering and coastal management, such as wave run-up and overtopping, the stability of rock and concrete structures, wave impacts and loadings, toe scour, and more.

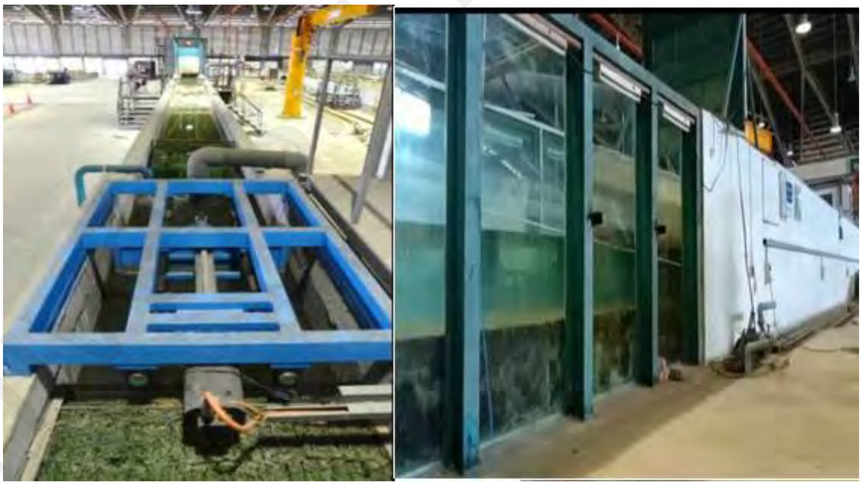


Figure 3.7: Wave flume at NAHRIM

3.3.2 Wave generator system

The wave generator system utilized in this experiment is the HR Merlin (previously known as HR Wavemaker) wave generation package, developed by HR Wallingford Ltd as shown in Figure 3.8. The purpose of this program is to control single and multi-element wavemakers to simulate various sea states in wave basins, wave flumes, and towing tanks. The program has the ability to generate both regular (sinusoidal) waves and random waves for various commonly used wave spectra using digitally filtered white noise. The facility also incorporates the capability to generate user-defined spectral shapes. For all tests in this project, irregular waves, such as random waves and natural sea waves are generated.

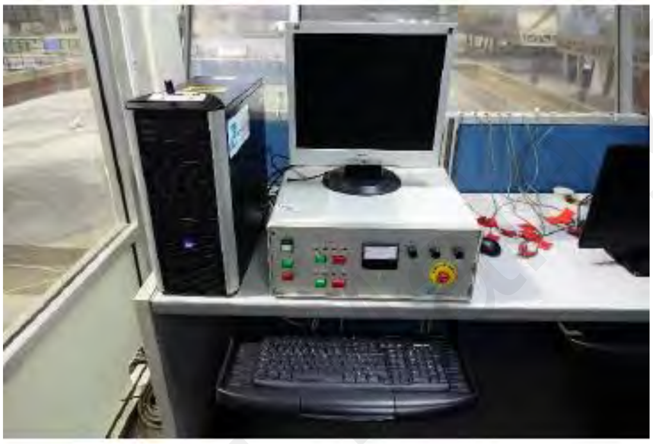


Figure 3.8: Wave generator

3.3.3 Data acquisition system

The Data Acquisition and Analysis Software Program, HR DAQ is capable to record 64 channels from a single analog-to-digital card, to utilize two eight-channel units with a frequency of 1000 Hz for a continuous 2 hours and to equate to 576 million data points (Figure 3.9). The term "wave counting" is not limited to water level recordings. Any output from an analog instrument, such as a displacement sensor, force transducer, or accelerometer, can be referred to as a "wave" record and analyzed as such. An analog signal appears as a wavy line, with crests and troughs appearing above and below the signal mean level.

HR DAQ utilizes a wave counting technique based on the upward crossings of the mean level of the signal. An upward crossing is detected when the previous value is below the crossing level and the present value is above or equal to it. This up-crossing technique defines a wave as occurring between two successive crossings, following international standards.

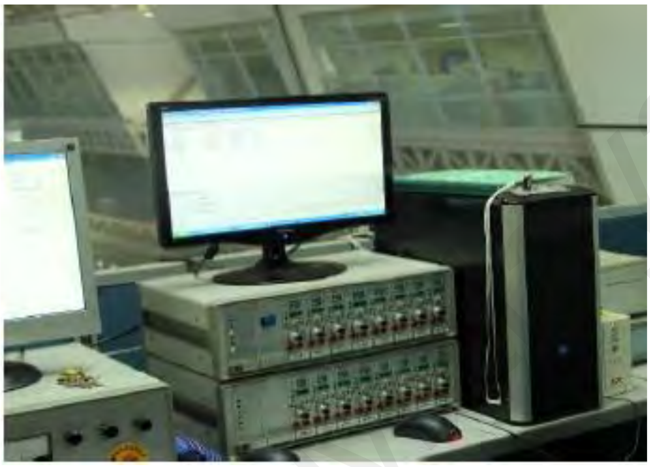


Figure 3.9: Data acquisition device

3.3.4 Experimental design and setup

Figure 3.10 shows a schematic diagram of the model setup. The slope of the structure used for this experiment is 1V:1H, and the crest width (G_c) is 0.5 m as illustrated in Figure 3.11. There are 5 wave probes (WP) utilized within the wave flume. Wave probes are instruments used to measure various properties of waves, such as wave height, wave period, and wave direction. There are several types of wave probes, including capacitive wave gauges, pressure sensors, and optical wave probes. Capacitive wave gauges measure the wave height by detecting the changes in capacitance caused by the movement of the water surface. Pressure sensors measure the pressure variations in the water caused by the passing waves. Optical wave probes use laser or photodiode technology to measure the wave height and other properties. Wave probes are typically mounted on a fixed structure or a floating platform and are used to obtain real-time or continuous measurements of the wave properties. The position of each wave prob plays a crucial role as it is the main source of collecting data such as the wave height and the wave period. For example, WP1 represents the first wave probe which is 12m away from the wave maker. The first wave probe allows for accurate measurement and characterization of the generated waves' properties, such as wave height, period, and velocity. This information is fundamental for understanding the input wave conditions and precisely controlling the wave generation process. Also, it acts as a quality control checkpoint, to verify that the generated waves meet the desired specifications. If any discrepancies are detected between the expected wave characteristics and the actual measurements, adjustments can be made to the wave generator settings before proceeding with the experiment. While WP 4 and WP5 are of great importance for wave overtopping study as the wave probe before the slope helps to observe and measure the waves characteristics as they propagate towards the slope. This information is crucial to understand wave transformation and potential changes in wave properties due to the presence of the slope. Moreover, the wave probe right before the structure allows for accurate detection of wave overtopping events. The amount of water spills over the structure during each wave event can be precisely quantified with the observation of the water level at this location. Also, Figure 3.12 illustrates the beach slope constructed in the wave flume with the core and underlayer.

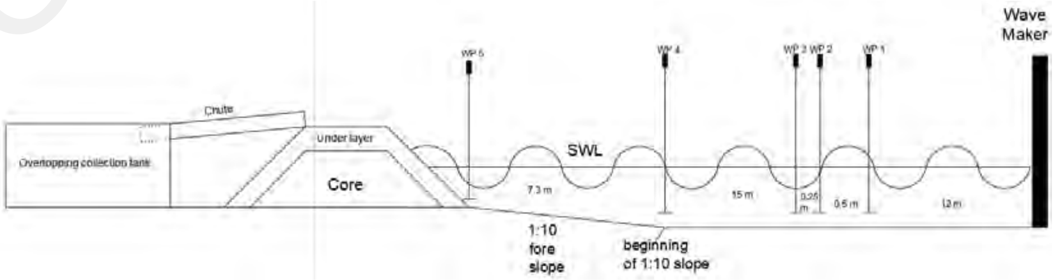


Figure 3.10: Experiment setup

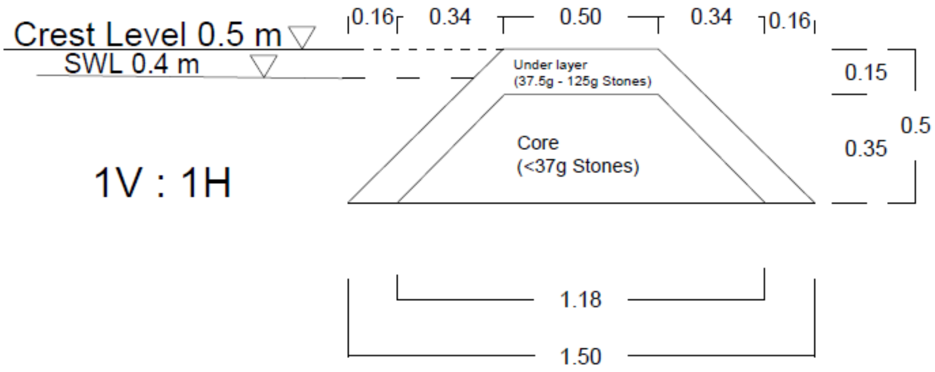


Figure 3.11: Beach structure layout

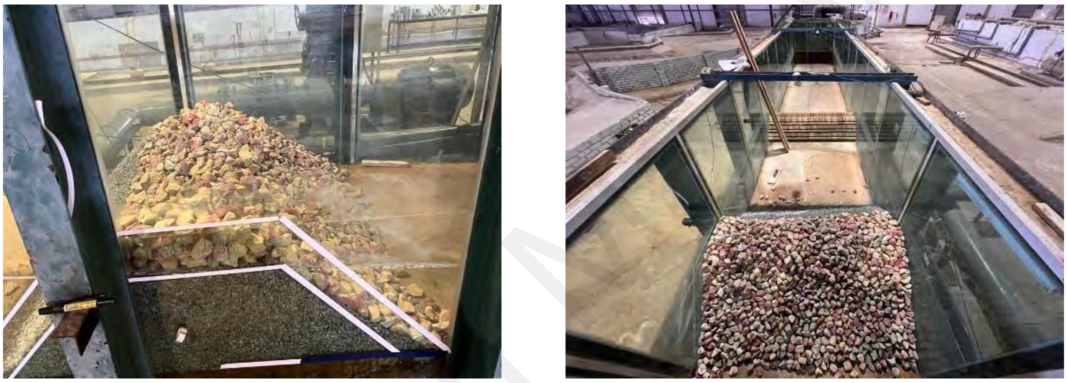


Figure 3.12: The beach structure in the wave flume

3.3.5 Wave conditions

The wave conditions are crucial for accurately measuring wave overtopping during a physical experiment. During this experiment, a total of 49 tests were conducted, each dependent on specific wave parameters such as wave height and wave period. These parameters align with the ones retrieved from the overtopping database (Eurotop, 2018). The primary objective is to examine the impact of the output data obtained from the physical experiment on the original database. The parameters considered for the physical experiment are summarized in Table 3.7.

Table 3.7: Summary of the parameters used in the experiment

#	Parameter	Unit	Definition
1	$H_{m0\ deep}$	m	Off-shore significant wave-height
2	$T_{m-1,\ deep}$	s	Off-shore spectral wave period
3	$h_{\ deep}$	m	Offshore water depth
4	h_t	m	water depth on the toe of a structure
5	$H_{m0\ t}$	m	Significant wave-height at the structure toe
6	$T_{p\ t}$	s	Peak wave period at the structure
7	CF		‘Complexity Factor’ The complexity of the test is indicated by a score with a possible range of 1 to 4.
8	RF		‘Reliability Factor’ The reliability of the test is indicated by a score with a possible range of 1 to 4.
9	q	$m^3/s/m$	mean wave overtopping discharge

The wave height refers to the vertical distance between the crest and trough of a wave. It is a crucial factor to determine the amount of overtopping that occurs, as higher waves are more likely to cause overtopping. The wave period, on the other hand, refers to the amount of time it takes for one complete wave cycle to pass a fixed point. Longer wave periods are generally associated with gentler waves, while shorter wave periods result in more energetic waves. Wavelength, which is the horizontal distance between two consecutive crests, is also a critical factor. A wave with a longer wavelength will generally have more energy and will therefore be more likely to cause overtopping. Moreover, Iribarren number, denoted as "ξ" or "Ir," is a dimensionless parameter used in coastal engineering to describe the stability of waves as they approach and interact with a sloping beach or coastal structure.

Porosity refers to the measure of void space or empty gaps within the layers of the structure. It is a fundamental parameter that characterizes the volume of open spaces or pores relative to the total volume of the material. Porosity plays a crucial role as the low value of porosity indicates that there are fewer void spaces within the layers of the structure, restricting the passage of water through them. This reduced permeability is essential in accurately gauging the amount of water that surpasses the coastal structure, which is the amount of wave overtopping. Below is the formula used for the wavelength (L), Iribarren number (ε) and porosity (n_{v}):

$$\boldsymbol{L = \frac{gT^2}{2\pi}} \tag{3.4}$$

$$\boldsymbol{\varepsilon = \frac{tan\alpha}{\sqrt{H/L}}} \tag{3.5}$$

$$\boldsymbol{n_v = 1 - \frac{\emptyset\, x\, D_n}{t_a}} \tag{3.6}$$

Finally, the slope of the structure being tested also plays a significant role in wave overtopping. A steep slope will result in a higher wave overtopping rate, while a more gradual slope will cause a lower rate of overtopping. However, only one slope was used in this experiment which is 1V:1H. Table 3.8 summarizes the wave conditions values that entered in the wave generator system as well as properties of the armour layer such as slope, porosity and Iribarren number.

Table 3.8: Theoretical wave conditions

#	H_{m0}	$T_{m-1,0}$	$L_{m-1,0}$	H/L	tan ß	IriBar
1	0.02	0.8	0.999745	0.020005	1	7.070167
2		1	1.562102	0.012803		8.837709
3		1.2	2.249427	0.008891		10.60525
4	0.06	0.8	0.999745	0.060015		4.081963
5		1	1.562102	0.03841		5.102454
6		1.2	2.249427	0.026673		6.122944
7	0.13	0.8	0.999745	0.130033		2.773148
8		1	1.562102	0.083221		3.466435
9		1.2	2.249427	0.057793		4.159721

3.3.6 Data collection

The data collection procedure involves a methodical approach of quantifying water volume through the use of a collection container, as visually depicted in Figure 3.13. This container captures the water overflow resulting from the experimental setup. Due to practical considerations, the width of the chute, set at 10 cm, serves as the primary means for channeling water into the container. To ascertain the precise water volume, the collected measurement must undergo division by 0.1, accounting for the specific width. Subsequently, this calculated volume is further divided by the total duration of the test to derive the wave overtopping value, q ($m^3/s/m$). This stepwise process ensures the accurate determination of wave overtopping rates and provides a comprehensive understanding of the observed phenomena within the experimental context.

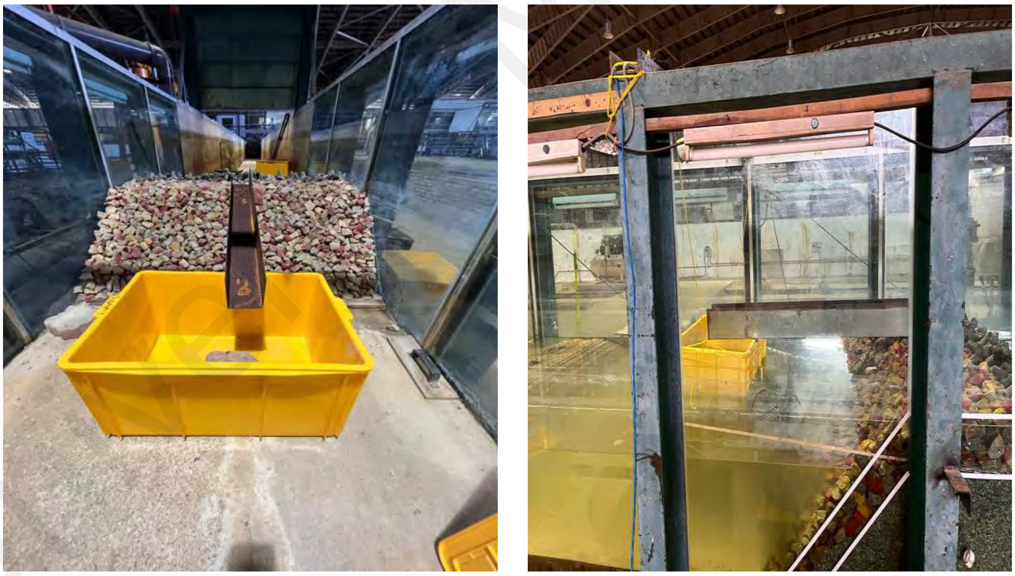


Figure 3.13: water collection container

3.4 Concluding remarks

The methodology chapter encompasses a thorough exploration of two crucial parts of the study, each contributing valuable insights to the study of wave overtopping. In the first part, a detailed overview of the dataset parameters was provided, with a keen focus on structural, hydraulic, and general parameters. Careful data exploration and preparation procedures were carried out to align with the study objectives, to ensure the dataset's suitability for model development. The importance of model tuning and scaling techniques was discussed in detail, emphasizing the significance of selecting appropriate hyperparameters and standardization for optimal results. The thoughtful selection and explanation of parameters for model training have set a strong foundation for the prediction model accuracy and robustness.

The second part of the study delves into the realm of physical experimentation, wherein a well-structured overview of the wave flume and its accompanying systems for wave generation and data collection was presented. Careful consideration of experimental design and setup underscores the attention to detail in the study, to ensure a controlled and reliable experimental environment. The importance of the slope type and position of wave probes highlights the significance of experimental conditions for accurate data collection. Moreover, a comprehensive presentation of the wave conditions and experimentally used parameters further enriches the experimental dataset.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

The Results and Discussion chapter marks the findings of the study of wave overtopping phenomena. This chapter presents a detailed account of the outcomes from both parts of the study, the predictive modeling and the physical experimentation conducted in NAHRIM. The findings from part one was utilized to discuss the performance and accuracy of the developed prediction model based on the XGBoost technique with using 30% of the database for testing. The results provided insight into various error indicators and the performance of the model across different q ranges. Also, the performance of the model was examined through the application of the bootstrap resampling method and the conduction of a comparison with an existing XGB model to validate the new model.

Part two of the study shows the experimental output acquired from the physical experiment. A total of 49 tests were conducted to collect the wave overtopping data. The details of the obtained wave conditions with the effect of the beach structure and the experiment setup were addressed. The results collected from the experiment were employed to create additional tests to be added to the original database. The outcome of adding these tests shows the impact of adding new experimental tests with different wave conditions to the original dataset.

4.2 Development of prediction model

4.2.1 XGB model performance

This section explores various performance metrics and presents a detailed discussion of the model outcomes. The range of the output parameter (q) used is discussed in section 3.2.4. The performance of the new model is assessed using a range of established evaluation measures and criterion. These criterion include error indicators such as Root Mean Square Error (RMSE), Bias, correlation of determination (R^2), Pearson correlation Coefficient (R), and the percentage error (PE). Below are the equations for each error indicator.

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x})}{N}}} \tag{4.1}$$

$$\boldsymbol{BIAS = \frac{1}{m}\sum_{i=1}^{m}\left(logq_{predicted} - logq_{measured}\right)} \tag{4.2}$$

$$\boldsymbol{R^2 = \frac{(\sum_{i=1}^{m}(logq_{measured} - logq_{measured})\ (logq_{estimated} - logq_{estimated}))^2}{\sum_{i=1}^{m}(logq_{measured} - logq_{measured})^2 \sum_{i=1}^{n}(logq_{estimated} - logq_{estimated})^2}} \tag{4.3}$$

$$\boldsymbol{R = \frac{n\ \Sigma XY - \Sigma X\ x\ \Sigma Y}{\sqrt{(n\Sigma X^2 - (\Sigma X)^2) x (n\Sigma Y^2 - (\Sigma Y)^2)}}} \tag{4.4}$$

$$\boldsymbol{Percentage\ error = \frac{|measured\ value - predicted\ value|}{meaasured\ value}\ x\ 100} \tag{4.5}$$

Table 4.1 presents the model performance for full database, training dataset and test dataset using error indicators. The RMSE value of 0.28 $m^3/s/m$ indicates a reasonable average deviation between the predicted and observed values in the test model. The Bias value of 0.002 suggests a negligible systematic deviation and unbiased model predictions. whereas the R^2 value of 0.91 and the R value of 0.95 signify a good fit to the test data, indicating a significant portion of the target variable variation is captured by the model. Although the error values are slightly higher compared to the Training model and Full model, the test model still yields a reasonable level of accuracy and precision. i.e., PE of 4.9%.

Table 4.1: Different error indicators for the datasets

Dataset	RMSE ($m^3/s/m$)	Bias	R^2	R	PE (%)
Full Database	0.15	0.005	0.96	0.98	2.2
Training dataset	0.07	0.007	0.99	0.99	1.02
Test dataset	0.28	0.002	0.91	0.95	4.9

It is also important to note that due to the application of standardization on the parameters, the RMSE method cannot be the sole governing error indicator for the model. Instead, the percentage error will be considered as the main metric for assessing the model performance in this thesis. The percentage error serves as a reliable indicator in scenarios where the magnitudes of the predicted and observed values may vary significantly. By expressing the prediction error as a percentage of the observed value, it provides a normalized measure that can be more meaningful and comparable across different scales. The percentage error of the model is measured at 4.9%, corresponding to an accuracy range of approximately 95.1% to 105.1%. In this context, a low percentage error implies that, on average, the model predictions deviate by approximately 4.9% from the true observed values. This also shows that the model predictions are, on average, within a reasonable margin of the actual values, with a relatively small percentage of error.

The percentage error takes into account the relative magnitude of the discrepancy between the predicted and observed values, rather than solely focusing on the absolute difference. This characteristic makes it particularly useful in situations where the target variable exhibits substantial variations in scale or where different observations have inherently different ranges. Such feature will ensure that the assessment of the model performance is not biased towards larger or smaller values. Instead, it emphasizes the relative accuracy of the predictions and provides a fair representation of the model ability to capture the underlying patterns and trends in the data. Thus, low percentage error indicates a reasonably accurate model performance, with predictions that are within an acceptable range of deviation from the observed values.

However, in extreme situations, such as small wave overtopping discharges, even relatively small errors, when compared to the measurement accuracy, may appear disproportionately large in terms of percentage deviation. While it is true that percentage errors can be misleading in certain scenarios, the achieved result observed here remains well within an acceptable range. This suggests that the model predictions align closely with the actual measurements, even when accounting for the inherent limitations and potential magnification of errors in percentage terms. Therefore, the obtained percentage error of 4.9% serves as a strong indication of the model capability to accurately estimate wave overtopping discharges in wave flumes. It highlights the model ability to capture the underlying dynamics and replicate the observed patterns with a commendable level of precision. The results obtained in Table 4.1 show a promising model with a low percentage error value (4.9%) and R value of 0.95 which indicates a high degree of correlation between the predicted and the actual values.

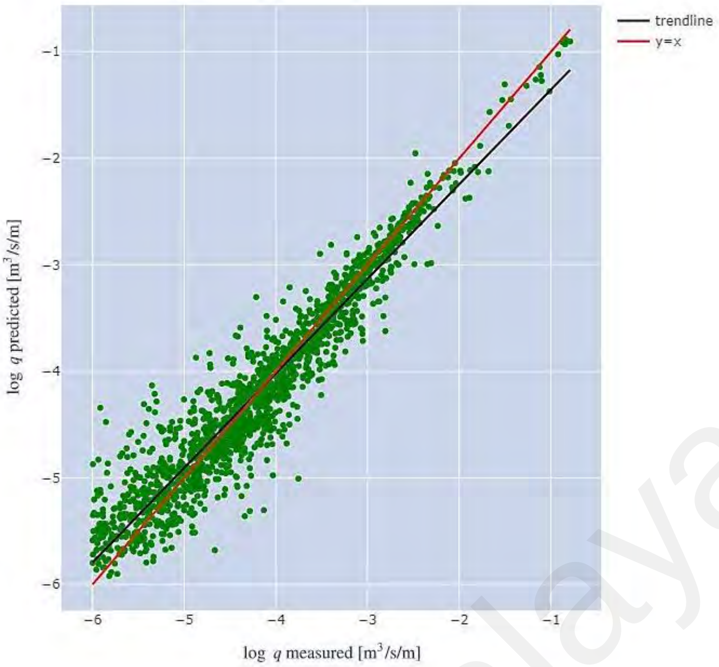


Figure 4.1: Performance of the XGB model (predicted vs measured) for test dataset (using log value).

Based on Figure 4.1, it can be observed that majority of data points clustered closely around the diagonal line, indicating a strong correlation between the predicted and observed values. However, some scattered points deviate slightly from the line, suggesting some level of prediction error. Overall, the figure demonstrates a reasonable agreement between the predicted and observed values. Figure 4.2 and Figure 4.3 illustrate about the performance of the training and full database respectively. The training database represents the data on which the model has been extensively trained, allowing it to learn and internalize the underlying patterns and relationships present in the training data. Therefore, it is natural and common to achieve a good of agreement between the predicted and measured values for the training database (Figure 4.2). The model has effectively utilized the information from the training data to make accurate predictions, resulting in a clustering of data points near the linear line. Similar results can be expected for the full database (Figure 4.3), which includes both the training and test datasets. This can be attributed to the fact that the full database encompasses all available information, allowing the model to leverage its inputs from the training data to make predictions on unseen instances.

The high level of agreement in both the training and full databases underscores the robustness and reliability of the model predictions. Overall perspective of the prediction error of the model can be seen in Figure 4.1 to Figure 4.3 as discussed earlier in Table 4.1.

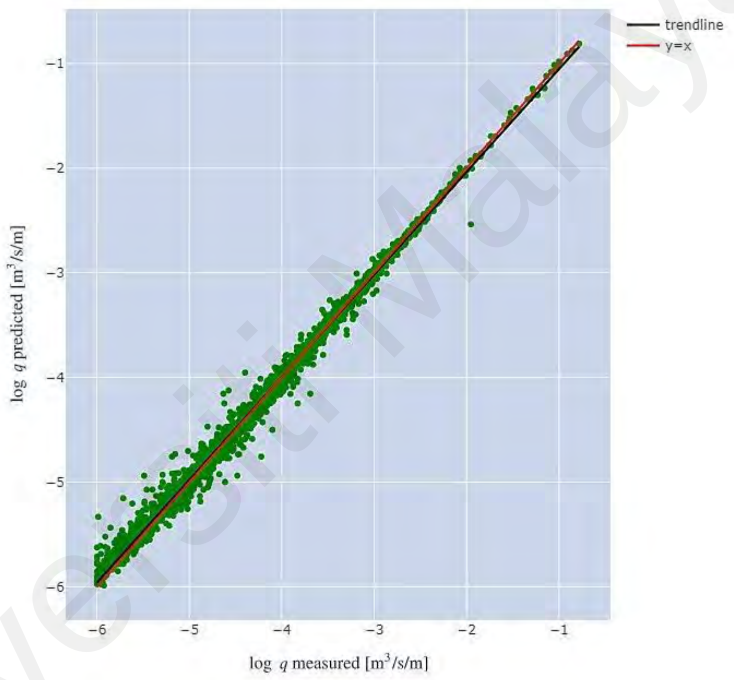


Figure 4.2: Performance of the XGB model (predicted vs measured) for training database (using log value).

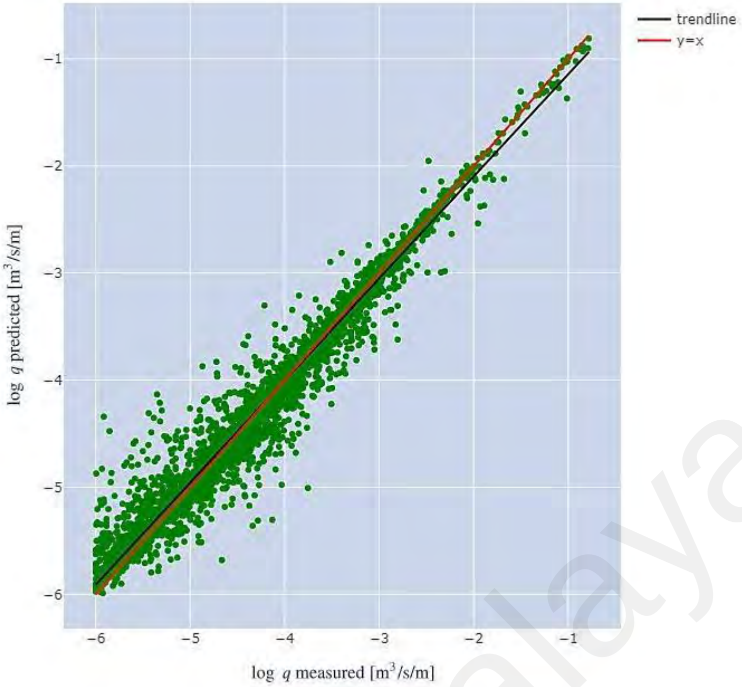
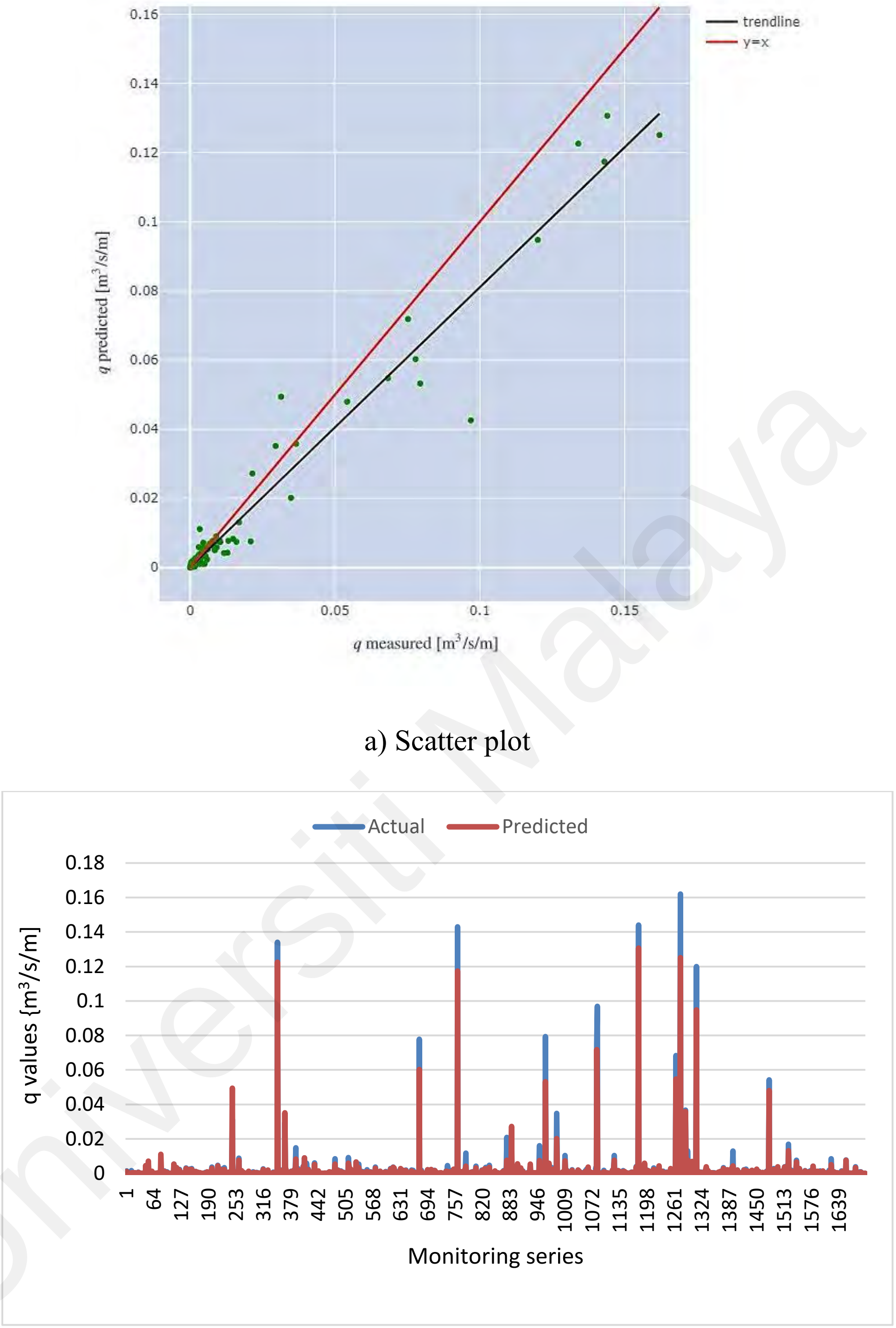


Figure 4.3: Performance of the XGB model (predicted vs measured) for full database (using log value)

Figure 4.4 presents model performance using the reversed values of q (actual values) based on scatter plot (Figure 4.4a)) and line graph (Figure 4.4b)), , to provide direct comparison with the log-scaled findings in Figure 4.1 to Figure 4.3. This comparison allows us to observe any differences, trends, or patterns that may have been influenced by the log transformation during preprocessing. Hence, by incorporating both log-scaled and actual values analyses in both preprocessing and post-processing stages, any bias and inconsistency issue in the model have been addressed. The figure yields a strong accuracy between the predicted and actual values. The good agreement between predicted and actual values also instils confidence in the model reliability and its capacity to make accurate predictions across various scenarios, despite the model underpredicted larger q values. Nevertheless, it suggests that the model has effectively captured the essential features of the data, avoiding overfitting or underfitting issues that could compromise its predictive tool.



a) Scatter plot

b) Line graph

Figure 4.4: XGB performance with the actual values of q (Actual vs Predicted)

4.2.2 XGB performance with different q ranges

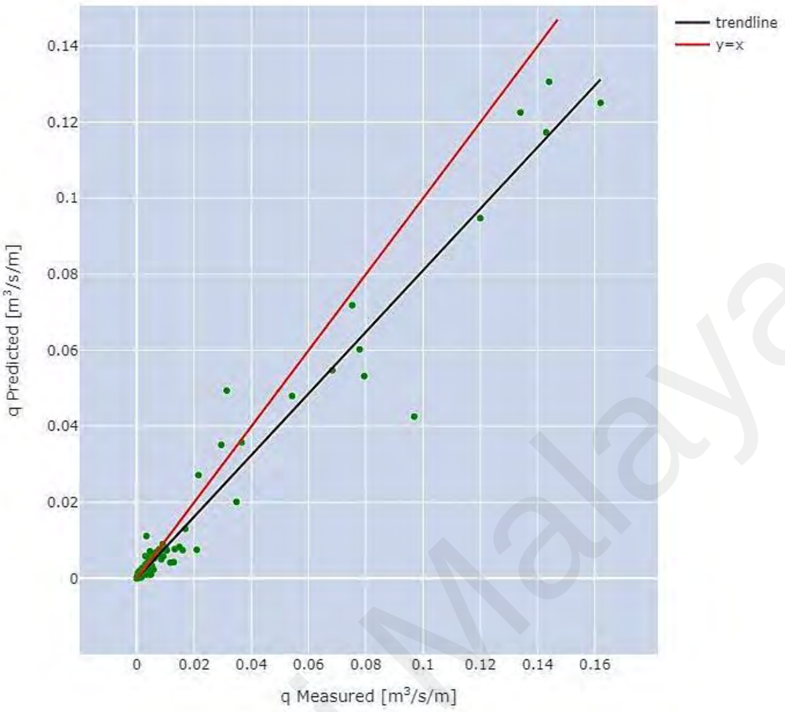
To further investigate the maximum capability and the performance of the new model for different ranges of q values, the test dataset containing 1701 data points are divided into three ranges. The data ranges are defined as follows: Low ($q < \mu - 2\sigma$), Medium ($\mu - 2\sigma \leq q \leq \mu + 2\sigma$), and High ($q > \mu + 2\sigma$), where q is the value of each data point, μ is the mean, and σ is the standard deviation of the entire dataset. Based on the previous studies by van Gent et al., (2007), it is apparent that an evaluation based on the relative error over the entire q range could cause a significant weight bias, mainly for low q values. Therefore, a technique called "partitioning analysis" is employed to precisely evaluate the performance of the model and assess the model capacity for each range (Nayak et al., 2005). Table 4.2 shows the results and the data points carried out for each range.

Table 4.2: Summary of the q range classification

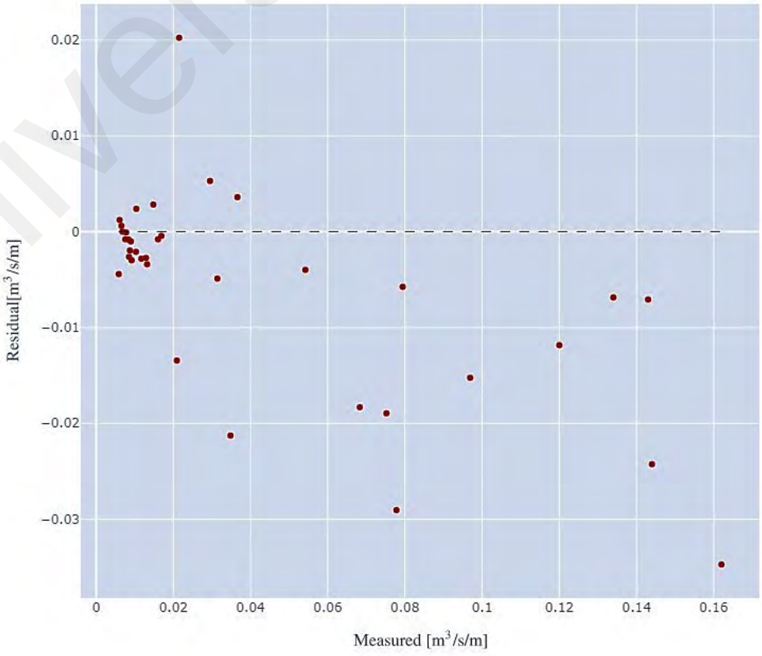
q range	Data points	RMSE ($m^3/s/m$)	R	PE (%)
Low	893	0.34	0.70	4.9
Medium	772	0.23	0.90	4.9
High	36	0.17	0.95	7.4

The model performance across different data ranges offers valuable insights into the model adaptability and predictive capabilities for varying subsets of the test database. Notably, the model demonstrates good degree of accuracy within the High range, consisting of 36 data points. This is reflected in the low RMSE value of 0.17 m3/s/m and R value of 0.95. The strong linear relationship between the model predictions and actual values in the high range highlights its proficiency in capturing underlying patterns in extreme data (Figure 4.5). In the medium range, encompassing 772 data points, the model continues to exhibit favourable performance. It achieved a competitive RMSE of 0.23

m^3/s/m and an acceptable R value of 0.90, signifying a robust linear association with the actual values as shown in Figure 4.6. This capability is attributed to the model ability to effectively accommodate data points clustered around the mean, thereby offering reliable predictions in this range. However, in the low range, comprising 893 data points, the model encounters slightly high errors. It exhibits an RMSE of 0.34 m^3/s/m and R value of 0.70, indicating a relatively weak linear relationship with the actual values as illustrated in Figure 4.7. Although the model can handle moderate variations in the medium range, the low range data presents challenge due to the dispersion of data points and potential outliers, consistent with the works already been done by (Zanuttigh et al., 2016a) and Formentin et al., (2017). The finding yields the effect of data lower than 10^{-6} m^3/s/m towards the model performance; thus, elimination of low range is needed. To fully address this issue, it requires further exploration and adjustments to the model architecture, whereby the treatment of outliers can be optimized to enhance performance. Moreover, the results indicate an unexpected pattern in the Percentage Error values, which do not align with the same sequence as the other error measurements, i.e., the low range exhibits a Percentage Error of 4.9% whereas the high range surprisingly shows a higher Percentage Error of 7.4%. This could be due to the imbalanced and uneven distribution of observations between different target classes in the training dataset. The imbalanced training data might have impacted the performance evaluation when subgrouping the database into low, medium, and high ranges using the Percentage Error metric. This discrepancy in performance could be due to the fact that the trained number of samples in the low range dataset significantly outweighs the number of samples in the high and medium ranges. Specifically, the high range dataset constitutes only 6.7% of the total training data, while the medium and low range dataset accommodate 39.27% and 54%, respectively. Nevertheless, it is evident that the model behaves robustly in all the ranges as the maximum value of percentage error is 7.4% which is considered as an acceptable value.



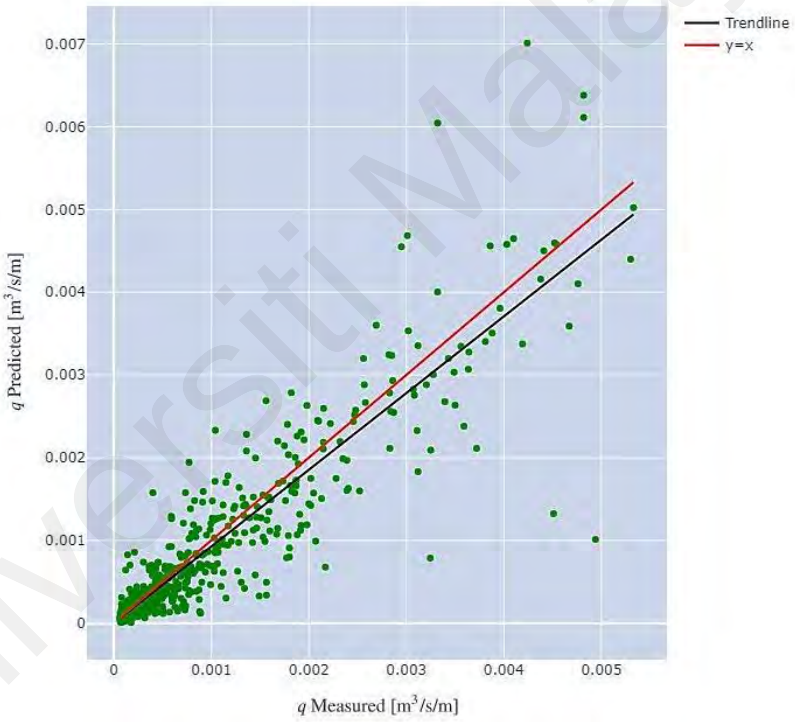
a)



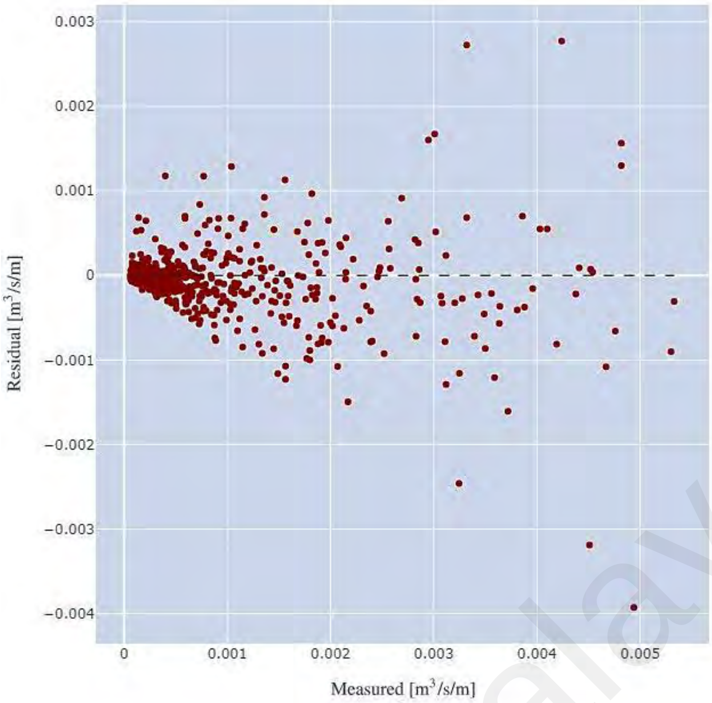
b)

Figure 4.5: Model performance for q high range

In Figure 4.5 (a), despite the fact that the number of points is too low, majority of the data points cluster tightly around the y=x line, demonstrating a satisfactory correlation. Only a few outliers exist, but they do not significantly affect the overall trend. This strong correlation signifies a robust relationship between the actual and predicted values, leading to an accurate model, within the high range. Whereas Figure 4.5 (b) shows the residuals, which are the differences between the predicted and actual values. The points are well scattered randomly around the residual line which indicates that the model does not exhibit any systematic bias in its predictions.



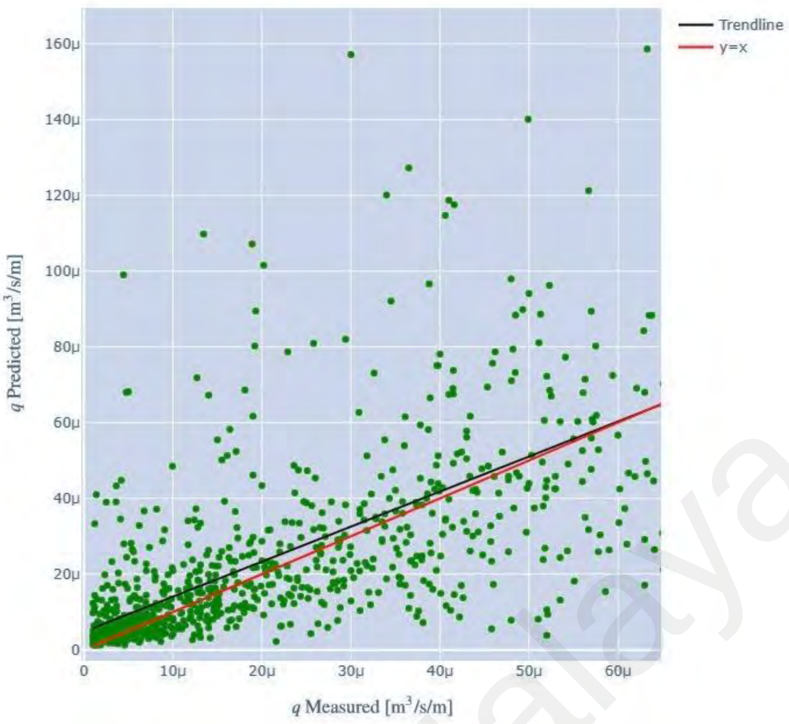
a)



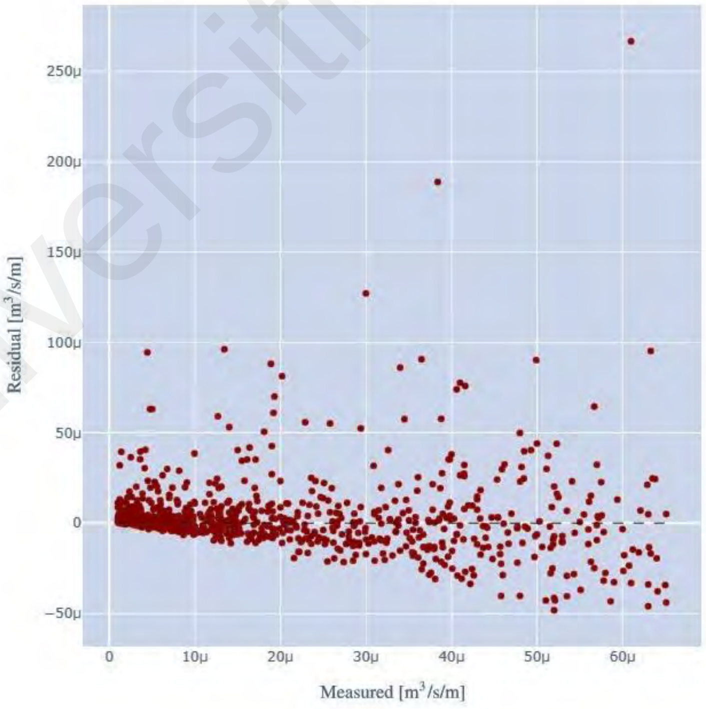
b)

Figure 4.6: Model performance for q medium range

For the medium range, Figure 4.6 (a) shows a consistent pattern within the data despite some scatters along the linear line. These scatters can be seen more clearly in tr the residual plot (Figure 4.6 (b)). There is a positive trend in the residuals, which means that the residuals are generally positive. This indicates that the model is overpredicted the overtopping as some outliers are detected within the dataset where the model yields poor prediction values. These outliers are positioned at higher q values within the dataset. There are also some outliers, which could indicate that there are some data points that are not well-fit by the model.



a)



b)

Figure 4.7: Model performance for q low range

The scatter plot (Figure 4.7 (a)) for low range shows a poor performance as the predicted values consistently surpass the actual values. This indicates a significant overestimation of overtopping by the predictive model in this range. Also, the presence of numerous outliers further emphasizes the lack of accuracy for this model. Moreover, in Figure 4.7 (b) there is abnormal behavior in the residuals displaying a distinct pattern. This pattern indicates that the presence of heteroscedasticity can be detected in the model. Therefore, subgrouping the database provides valuable insights into the performance of the three ranges and the impact of these ranges on the model performance. The observed performance for the low range comes as no surprise, since it has been stated by previous studies. However, by subgrouping the database in this study, the effect of this specific range becomes more apparent and visible.

4.2.3 Bootstrap resampling method

As part of the model validation, bootstrap resampling method is applied to obtain estimates for the uncertainties in the model predictions. This approach is summarized as follows. From the total training data 500 bootstrap resamples are generated. Each bootstrap resample forms the training set for a new model and contains a different randomized selection from the total data set. In that training, the data points not selected in the resample serve as the validation dataset. This validation set is used for ‘early stopping’, which stops the training of a single model after 1000 consecutive additional trees fail to improve the model performance on the validation data set. The data in the validation set also changes with each resample. In the end, this gives an ensemble of 500 retrained (’resampled’) models, where all suitable data is utilized while no single model is trained on the entire data set. From the outcome of the 500 models, it is found that the mean and the median values are not the best prediction of all outcomes, as shown in Table 4.3. It seems that the tails of the distribution of the bootstrapped model predictions are not symmetrical.

The insights extracted from the bootstrap results show that the performance of the model is better without applying the bootstrap method. This is demonstrated in Table 4.3 where error indicator values associated with the finest model generated through bootstrap resampling are comparatively higher than those of the model without such resampling, see Table 4.1. This finding highlights an unexpected outcome where the conventional approach outperforms the bootstrap resampling method. However, after analyzing the best resampled model among the 500 resampled models, it was found that the number of tests in the high and low q range is 35 and 3466 tests respectively. This is in contrast to the high q range for the original training dataset, which encompasses 265 and 2146 tests respectively. The imbalance in training data, particularly in the low q range, is known to negatively affect model performance, as discussed, and shown in section 4.2.2.

Table 4.3: Performance of the model with bootstrap method

	RMSE ($m^3/s/m$)	R	PE (%)
Mean value	0.329	0.941	5.7
Median value	0.328	0.942	5.7
Best value	0.311	0.947	5.4

Additionally, a confidence interval can also be determined from the large number of model predictions. A confidence interval in bootstrap resampling is constructed by taking percentiles from this distribution. For example, a 95% confidence interval would involve calculating the 2.5th percentile and the 97.5th percentile of the distribution. The range between these two percentiles represents the confidence interval. This indicates that through multiple repetitions of the resampling procedure, the actual parameter of interest should approximately fall within the interval 95% of the time. Another observation is that the confidence interval derived with the bootstrap resampling method is very small. As can be seen Figure 4.8, the lines representing the upper and lower limits on a graph are very close to the mean which indicates that there is a high level of confidence in the accuracy and precision of the predicted values. Also, it suggests that the model is consistently providing a reliable prediction value, and there is minimal room for the actual values to deviate from this estimate.

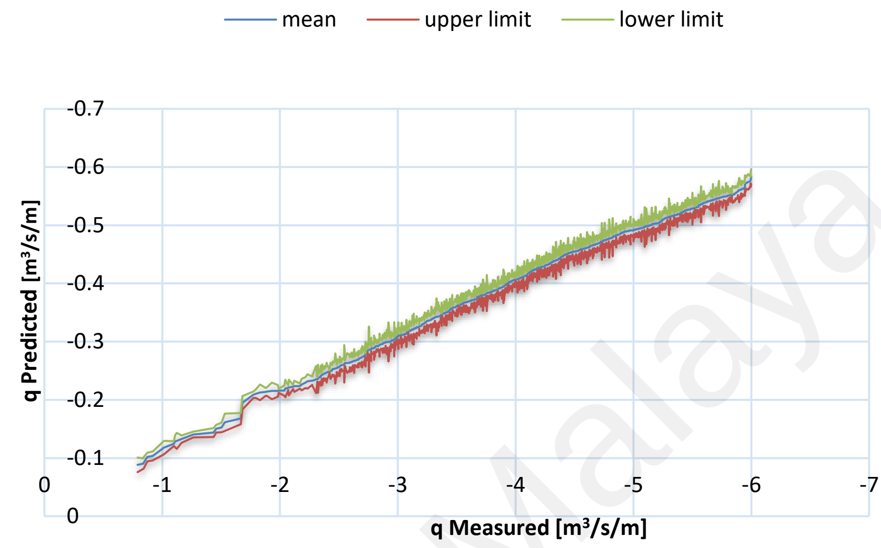


Figure 4.8: wave overtopping predictions versus measured wave overtopping, including 95% confidence interval from bootstrap resampling.

4.2.4 Comparison with existing XGB model

The performance of the new model is compared with an established XGB model introduced by den Bieman et al., (2021), henceforth DB., with different filtration processes, WF values, and more hyperparameter values as described in section 3.2.5.1. Table 4.4 shows the performance indicator of the training dataset, full database and test dataset for both models. Note that the only error indicator that can be directly compared with the DB model is the RMSE method since it is the sole indicator utilized in both studies. The table shows that there is a slight improvement in performance observed for both the full and test databases. The new study achieved RMSE values of 0.150 m^3/s/m and 0.280 m^3/s/m for the full and test databases, respectively, while the existing study obtained RMSE values of 0.154 m^3/s/m and 0.284 m^3/s/m for the same database. Although the new study demonstrates lower RMSE values, the difference between the results of both studies is not significant, indicating a modest disparity in model performance. When considering the full database, the new study RMSE value of 0.150 m^3/s/m is marginally better than the DB model with RMSE value of 0.154 m^3/s/m respectively. Similarly, for the test database, the new mode yields RMSE value of 0.280 m^3/s/m, a slight improvement compared to the DB with the RMSE value of 0.284 m^3/s/m respectively. These small differences suggest that the model of the new study captures the underlying patterns and relationships in the data slightly more accurately, resulting in better performance than DB model.

Table 4.4: Comparison between New and existing XGB model by (den Bieman et al., 2021) in terms of RMSE

Database	New XGB	DB
Full database	0.150	0.154
Training dataset	0.070	0.098
Test dataset	0.280	0.284

In addition to the performance on the full and test databases, there is a notable difference in the RMSE values between the two models in the training dataset. The DB model obtained RMSE value of 0.098 m^3/s/m, while the new model achieved a lower RMSE value of 0.07 m^3/s/m. This discrepancy in the RMSE values for the training dataset suggests that the new model has a better fit to the training data compared to the DB model. It also indicates that the new model has good predictions, on average, to the actual values within the training dataset and the model is better trained to minimize the errors between the predicted and actual values.

4.3 Physical experiment

4.3.1 Wave conditions values

The measured values from the physical experiment are displayed in Table 4.5. The values detected from the wave probes may differ from the ones entered in the system due to various factors and sources of error. Firstly, measurement error is a common issue as wave probes are sensitive instruments, and even minor misalignments or calibration errors can lead to inaccuracies. Wave reflections and refractions within the flume can alter the wave patterns and cause differences between the generated wave values and the measured values at specific locations. Secondly, wave absorption along the flume walls or other materials can lead to energy loss and changes in wave characteristics from the initial input values. Boundary effects near the flume walls might also influence the wave behavior and impact the measured values. Furthermore, the sensitivity levels and operating ranges of different wave probes can vary and introduce differences in the measurements. The dynamic nature of the wave flume environment, with changing temperature, pressure, and water flow, can also contribute to the differences in detected values. Lastly, flow interference and turbulence around structures can affect the accuracy of the wave probe measurements.

Table 4.5: Actual wave conditions

#	H_{m0}	$T_{m-1,0}$	$L_{m-1,0}$	H/L
1	0.16	1.2	22.17	0.0072
2		1.3	26.02	0.0061
3		1.4	30.18	0.0053
4	0.17	1.2	22.17	0.0076
5		1.3	26.02	0.0065
6		1.4	30.18	0.0056
7	0.2	1.2	22.17	0.0090
8		1.3	26.02	0.0076
9		1.4	30.18	0.0066

4.3.2 Performance of the new tests

To properly assess the performance of the model using the new tests, two models were developed with the same parameters of the existing dataset as presented in Table 3.7. The first model relied only on the existing dataset, whereas the second model incorporated the new tests into the existing database. The impact of the new tests on the existing dataset is evaluated using some error evaluation indicators such as percentage error and coefficient of correlation as stated in Table 4.6. Also, scatter diagrams of the actual values versus the predicted values for both old dataset (original) and new dataset (mix) is shown in Figure 4.9.

Table 4.6: Error indicators for original dataset and mix dataset

	RMSE ($m^3/s/m$)	R^2	R	PE (%)
Original dataset	0.55	0.69	0.83	10.09
Mix dataset	0.56	0.66	0.81	10.43

The table shows that the values are relatively close to each other with slightly better results for the original dataset. This implies a weak effect of the new tests on the original tests. Such findings can be due to the complexity of the model setup in the new dataset and the physical laboratory settings. The new data were designed using multiple types of rubble mound structures that yield roughness factor of less than 1 (simple impermeable slope/structure). This circumstance is not being well represented in the training data of the model.

Also, another element that could affect is the wave reflection. Wave reflection occurs when incident waves are bounced back upon encountering a boundary, which can distort the wave pattern and alter the overall wave characteristics. As can be seen in Figure 4.10, the reflection coefficients of these tests are ranges between 0.48 and 0.5 which means that there is a significant effect on the incident waves.

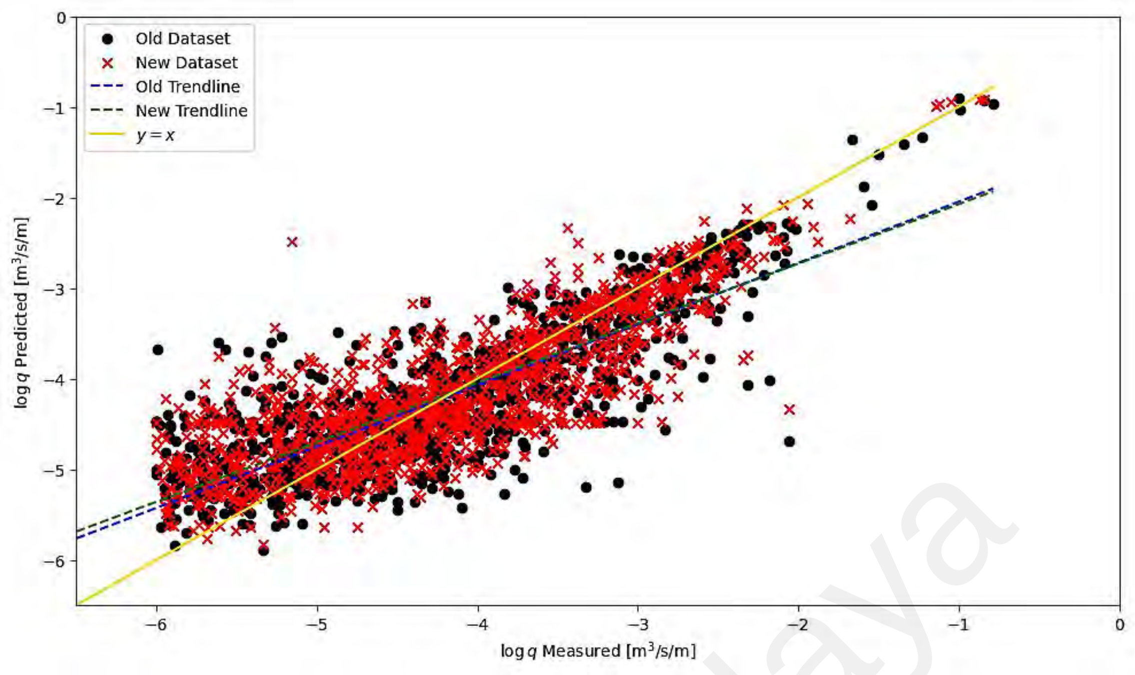


Figure 4.9: Comparison between old dataset and new dataset (using log values)

The alignment of most scatter plot points underscores a substantial correlation between the two datasets, indicating that the introduction of new tests has an impact. Most points are clustered around a central region, where predicted values closely match their actual counterparts. This highlights the efficacy of predictive models for both old and new tests. It also gives us a clear picture of the impact of new tests on the original dataset.

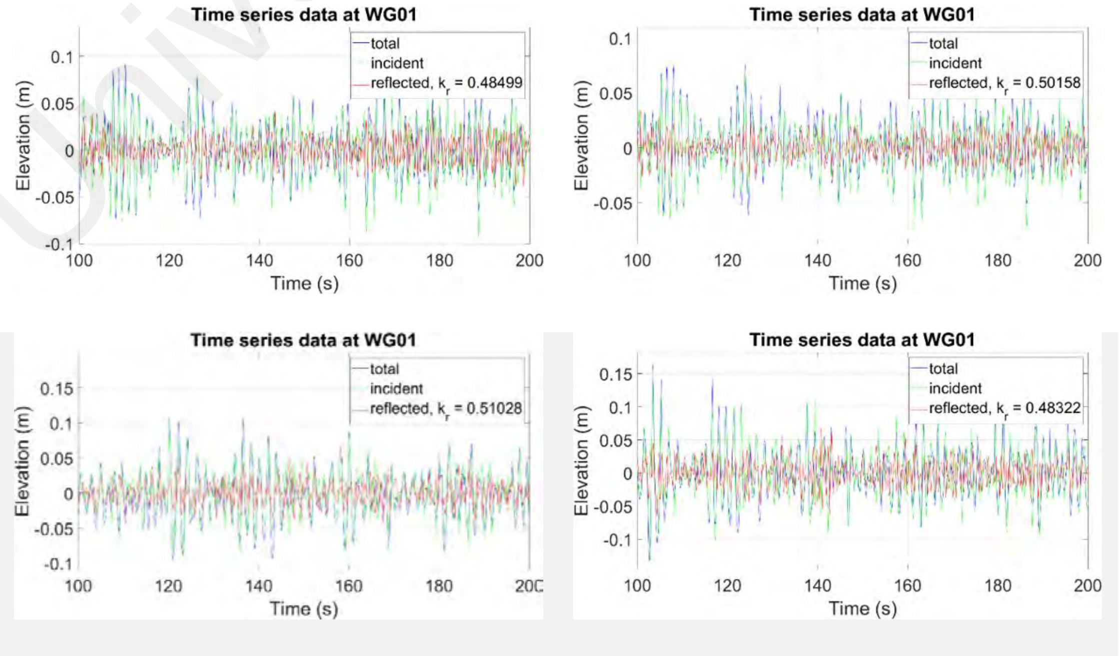


Figure 4.10: Surface elevation time series with reflection coefficient data

The existence of wave reflection has a significant influence on incident waves, causing notable changes in their behavior and attributes which can be seen from the figure. The consistent elevation of the total wave, being notably higher than that of the incident wave across various time points conveys the profound influence of the reflected wave on the incident wave. The reflection coefficients vary from 0.48 to 0.5 implies a substantial reflection component. This suggests that a significant portion of the wave energy is being reflected. The reflected energy combines with the incident wave energy to form the total wave. Consequently, the elevation of the total wave exceeds that of the incident wave, and this difference is consistently evident throughout the observed time span. This elevation disparity holds implications for wave dynamics. The reflected wave, upon encountering the boundary or interface, interacts with the incident wave, resulting in constructive interference. This interaction contributes additional energy to the total wave, elevating its overall amplitude and effecting the outcome of the experiment.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

This thesis has presented an investigation into the prediction of wave overtopping with a focus on developing a new model using the XGB algorithm. The overtopping process with the parameters of the database has been studied in detail by giving a step-by-step procedure for the development of the XGB model. The different ranges of the wave overtopping data have been explored by studying the performance of each range. A physical experiment has been conducted and the data has been added to the existing database to assess the performance before and after adding the new tests. The different ranges of the wave overtopping data have been explored by studying the performance of the model in each range. This chapter summarizes the findings and highlights the important outcomes of this research.

5.1 Research Conclusion

In conclusion, this thesis has undertaken a comprehensive exploration of wave overtopping prediction, employing the robust XGBoost (XGB) algorithm. The research encompassed two parts: the development of a predictive model with validating its predictions findings and conducting a physical experiment to collect overtopping data.

The first part of the study delved into the details of model development. The foundation for accurate predictions was laid by analyzing the database, preprocessing the data, and using different hyperparameters. Utilizing hyperparameters and fine-tuning process showed the potential of XGB algorithm in capturing wave overtopping. The model is reliable with the RMSE values of 0.28 $m^3/s/m$. The model also performs well in the prediction of different ranges of q with RMSE values of 0.34 $m^3/s/m$, 0.23 $m^3/s/m$, and 0.17 $m^3/s/m$ for low, medium, and high ranges respectively. Similarly, the percentage error statistics of 4.9%, 4.9%, and 7.4% for these ranges and 4.9% overall, reflect the model capability in quantifying discrepancies. The high range had a higher percentage error because the data is evenly distributed. The preference for this metric was rationalized for its comprehensive representation of prediction accuracy. The model categorization into ranges enhances our understanding of its performance in diverse situations. Model validation, employing bootstrap resampling, helped in checking the reliability of the model. Remarkably, the model inherent strength was validated as it outperformed the resampling-based approach. As a part of the validation, the new model is compared to an existing model by DB and the new model outperformed the existing model, proving the effectiveness of the XGB technique and making a significant contribution to the field.

The second phase of the thesis unfolded through physical experiment, introducing new empirical data to the existing database. The design and execution of the experiment within the National Hydraulic Research Institute of Malaysia (NAHRIM) underscored the practical complexities inherent in empirical data collection. The research was enriched by detailed information on the wave flume, generator system and data collection. However, adding only 49 experimental tests to the current database showed the difficulties of conducting enough tests to make a difference in a large dataset. Introducing the 49 new tests to the existing dataset did not result in significant differences in the performance of the model. The percentage error for the original dataset was 10.09% and 10.43% after the addition of the new tests, with the RMSE values of 0.55 m^3/s/m and 0.56 m^3/s/m, respectively. The slight drop in performance could be attributed to factors such as wave reflection, the sensitivity of the model to wave characteristics or the complexity of the model setup.

Collectively, this thesis offers a robust estimation of wave overtopping rates. Using the XGB algorithm with experimental data not only advances coastal engineering practices but also contributes to our knowledge about making predictions using different datasets.

5.2 Recommendation for future work

Despite the work done on the estimation of the wave overtopping using XGB algorithm and the physical experiment conducted to obtain new overtopping data, it is clear that there are important factors that are not fully covered, due to existing limitations. The Eurotop database has a lot of parameters with many tests that require more studying to fully acquire the maximum usage of this database. Further exploration is necessary to understand why the low range of the output parameter, q, negatively affects the performance of machine learning algorithms applied to the database. This way, there is no requirement to eliminate the low range from the database, and the potential exists to utilize more data for the improvement of machine learning models. More research should focus on fine-tuning the hyperparameters of the XGBoost algorithm, to determine the most suitable settings for specific datasets. Systematically evaluating various hyperparameter combinations will unlock the full potential of the algorithm and enhance its accuracy. Further analysis and examination of additional metrics or factors related to model performance should be conducted to gain a comprehensive understanding especially that RMSE might not be applicable in all cases, thus exploring alternative evaluation metrics can provide deeper insights. To ensure a comprehensive dataset, diverse test conditions encompassing a wide range of wave characteristics need to be incorporated. Advanced measurement techniques such as high-resolution cameras and laser systems must be employed for precise data collection. Scale and geometric similarity between the experimental setup and real coastal scenarios must be maintained, to enhance data applicability. To assess how the newly collected data affects the current database, it is advisable to conduct numerous tests—ideally involving more than half of the existing database. This approach ensures a clear understanding of the impact of the process.

Considering the acceptable performance of XGBoost model in the estimation of wave overtopping rates in this thesis, application of the model in other related studies can lead to significant contribution.

REFERENCES

Adnan, R. M., Liang, Z., El-Shafie, A., Zounemat-Kermani, M., & Kisi, O. (2019). Prediction of suspended sediment load using data-driven models. *Water (Switzerland)*, *11*(10). https://doi.org/10.3390/w11102060

Afan, H. A., El-shafie, A., Mohtar, W. H. M. W., & Yaseen, Z. M. (2016). Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. *Journal of Hydrology*, *541*, 902–913. https://doi.org/10.1016/j.jhydrol.2016.07.048

Ahmad, A., Razali, S. F. M., Mohamed, Z. S., & El-shafie, A. (2016). The Application of Artificial Bee Colony and Gravitational Search Algorithm in Reservoir Optimization. *Water Resources Management*, *30*(7), 2497–2516. https://doi.org/10.1007/s11269-016-1304-z

Allawi, M. F., Jaafar, O., Mohamad Hamzah, F., Abdullah, S. M. S., & El-shafie, A. (2018). Review on applications of artificial intelligence methods for dam and reservoir-hydro-environment models. *Environmental Science and Pollution Research*, *25*(14), 13446–13469. https://doi.org/10.1007/s11356-018-1867-8

Baldock, T. E., Hughes, M. G., Day, K., & Louys, J. (2005). Swash overtopping and sediment overwash on a truncated beach. *Coastal Engineering*, *52*(7), 633–645. https://doi.org/10.1016/j.coastaleng.2005.04.002

Baldock, T. E., Peiris, D., & Hogg, A. J. (2012). Overtopping of solitary waves and solitary bores on a plane beach. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *468*(2147), 3494–3516. https://doi.org/10.1098/rspa.2011.0729

Chen, S. H., Jakeman, A. J., & Norton, J. P. (2008). Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation*, *78*(2–3), 379–400. https://doi.org/10.1016/j.matcom.2008.01.028

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & He, T. (2014). xgboost: Extreme Gradient Boosting. *R Lecture*, *2016*, 1–84.

Chen, W., Marconi, A., van Gent, M. R. A., Warmink, J. J., & Hulscher, S. J. M. H. (2020). Experimental study on the influence of berms and roughness on wave overtopping at rock-armoured dikes. *Journal of Marine Science and Engineering*, *8*(6), 1–21. https://doi.org/10.3390/jmse8060446

Chini, N., & Stansby, P. K. (2012). Extreme values of coastal wave overtopping accounting for climate change and sea level rise. *Coastal Engineering*, *65*, 27–37. https://doi.org/10.1016/j.coastaleng.2012.02.009

De Rouck, J., Geeraerts, J., Troch, P., Kortenhaus, A., Pullen, T., & Franco, L. (2005). New results on scale effects for wave overtopping at coastal structures. *International Conference on Coastlines, Structures and Breakwaters 2005: Harmonising Scale and Detail - Proceedings of the International Conference on Coastlines, Structures and Breakwaters 2005*, *2006*(February), 29–43.

De Rouck, J., Van de Walle, B., VAN DAMME, L., WILLEMS, M., FRIGAARD, P., & MEDINA, J. (2001). Prototype measurements of wave run-up on a rubble mound breakwater. *Waves*, 1–14.

De Rouck, J., Verhaeghe, H., & Geeraerts, J. (2009). Crest level assessment of coastal structures - General overview. *Coastal Engineering*, *56*(2), 99–107. https://doi.org/10.1016/j.coastaleng.2008.03.014

Dehghani, A., Mohammad, H., Hiyat, Z., & Mortazavizadeh, F. (2023). Ecological Informatics Comparative evaluation of LSTM , CNN , and ConvLSTM for hourly short-term streamflow forecasting using deep learning approaches. *Ecological Informatics*, *75*(February), 102119. https://doi.org/10.1016/j.ecoinf.2023.102119

den Bieman, J. P., van Gent, M. R. A., & van den Boogaard, H. F. P. (2021). Wave overtopping predictions using an advanced machine learning technique. *Coastal Engineering*, *166*(November 2020), 103830. https://doi.org/10.1016/j.coastaleng.2020.103830

den Bieman, J. P., Wilms, J. M., van den Boogaard, H. F. P., & van Gent, M. R. A. (2020). Prediction of mean wave overtopping discharge using gradient boosting decision trees. *Water (Switzerland)*, *12*(6), 1–13. https://doi.org/10.3390/W12061703

Ding, Z., Nguyen, H., Bui, X. N., Zhou, J., & Moayedi, H. (2020). Computational Intelligence Model for Estimating Intensity of Blast-Induced Ground Vibration in a Mine Based on Imperialist Competitive and Extreme Gradient Boosting Algorithms. *Natural Resources Research*, *29*(2), 751–769. https://doi.org/10.1007/s11053-019-09548-8

Elizaga, N. B., Maravillas, E. A., & Gerardo, B. D. (2014). Regression-based inflow forecasting model using exponential smoothing time series and backpropagation methods for Angat Dam. *2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2014 - 7th HNICEM 2014 Joint with 6th*

International Symposium on Computational Intelligence and Intelligent In, *November 2013*. https://doi.org/10.1109/HNICEM.2014.7016185

Etemad-Shahidi, A., Bali, M., & van Gent, M. R. A. (2021). On the toe stability of rubble mound structures. *Coastal Engineering*, *164*(October 2020), 103835. https://doi.org/10.1016/j.coastaleng.2020.103835

Etemad-Shahidi, A., & Jafari, E. (2014). New formulae for prediction of wave overtopping at inclined structures with smooth impermeable surface. *Ocean Engineering*, *84*, 124–132. https://doi.org/10.1016/j.oceaneng.2014.04.011

Etemad-Shahidi, A., Shaeri, S., & Jafari, E. (2016). Prediction of wave overtopping at vertical structures. *Coastal Engineering*, *109*, 42–52. https://doi.org/10.1016/j.coastaleng.2015.12.001

Eurotop. (2016). *EurOtop: Manual on wave overtopping of sea defences and related structures*. 264. www.overtopping-manual.com

Eurotop. (2018). *Eurotop 2018; Manual on wave overtopping of sea defences and related structures. An overtopping manual largely based on European research, but for worldwide application*. 320. www.overtopping-manual.com

Filippo, A., Rebelo Torres, A., Kjerfve, B., & Monat, A. (2012). Application of Artificial Neural Network (ANN) to improve forecasting of sea level. *Ocean and Coastal Management*, *55*, 101–110. https://doi.org/10.1016/j.ocecoaman.2011.09.007

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), 1–31. https://doi.org/10.1142/S0578563417500061

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J., Tibshirani, R., & Hastie, T. (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016120463

Gallach-Sánchez, D., Troch, P., & Kortenhaus, A. (2021). A new average wave overtopping prediction formula with improved accuracy for smooth steep low-crested structures. *Coastal Engineering*, *163*(February 2020). https://doi.org/10.1016/j.coastaleng.2020.103800

Gallien, T. W., Sanders, B. F., & Flick, R. E. (2014). Urban coastal flood prediction: Integrating wave overtopping, flood defenses and drainage. *Coastal Engineering*, *91*, 18–28. https://doi.org/10.1016/j.coastaleng.2014.04.007

Geeraerts, J., Troch, P., De Rouck, J., Verhaeghe, H., & Bouma, J. J. (2007). Wave overtopping at coastal structures: prediction tools and related hazard analysis. *Journal of Cleaner Production*, *15*(16), 1514–1521. https://doi.org/10.1016/j.jclepro.2006.07.050

Goda, Y. (2009). Derivation of unified wave overtopping formulas for seawalls with smooth, impermeable surfaces based on selected CLASH datasets. *Coastal Engineering*, *56*(4), 385–399. https://doi.org/10.1016/j.coastaleng.2008.09.007

Hanoon, M. S., Ahmed, A. N., Fai, C. M., Birima, A. H., Razzaq, A., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2021). Application of Artificial Intelligence Models for modeling Water Quality in Groundwater: Comprehensive Review, Evaluation and Future Trends. *Water, Air, and Soil Pollution*, *232*(10).

https://doi.org/10.1007/s11270-021-05311-z

Hosseinzadeh, S., Etemad-Shahidi, A., & Koosheh, A. (2021). Prediction of mean wave overtopping at simple sloped breakwaters using kernel-based methods. *Journal of Hydroinformatics*, *23*(5), 1030–1049. https://doi.org/10.2166/hydro.2021.046

Ibrahim, D. (2016). An Overview of Soft Computing. *Procedia Computer Science*, *102*(August), 34–38. https://doi.org/10.1016/j.procs.2016.09.366

Ibrahim, K. S. M. H., Huang, Y. F., Ahmed, A. N., Koo, C. H., & El-Shafie, A. (2022). A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alexandria Engineering Journal*, *61*(1), 279–303. https://doi.org/10.1016/j.aej.2021.04.100

Ibrahim, M. S. I., & Baldock, T. E. (2020). Swash overtopping on plane beaches – Reconciling empirical and theoretical scaling laws using the volume flux. *Coastal Engineering*, *157*(November 2019), 103668. https://doi.org/10.1016/j.coastaleng.2020.103668

Ibrahim, M. S. I., & Baldock, T. E. (2021). Physical and Numerical Modeling of Wave-by-Wave Overtopping along a Truncated Plane Beach. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *147*(5), 1–18. https://doi.org/10.1061/(asce)ww.1943-5460.0000663

Ibrahim, R. K., Afan, H. A., El-shafie, A., & Fai, C. M. (2019). *Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios*.

Igboanugo, A. C. (2013). Predicting Water Levels at Kainji Dam Using Artificial Neural Networks. *Nigerian Journal of Technology*, *32*(1), 129-136–136.

Jumin, E., Zaini, N., Ahmed, A. N., Abdullah, S., Ismail, M., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2020). Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. *Engineering Applications of Computational Fluid Mechanics*, *14*(1), 713–725. https://doi.org/10.1080/19942060.2020.1758792

Koosheh, A., Etemad-Shahidi, A., Cartwright, N., Tomlinson, R., & Hosseinzadeh, S. (2020). the Comparison of Empirical Formulae for the Prediction of Mean Wave Overtopping Rate At Armored Sloped Structures. *Coastal Engineering Proceedings*, *36v*, 22. https://doi.org/10.9753/icce.v36v.structures.22

Koosheh, A., Etemad-Shahidi, A., Cartwright, N., Tomlinson, R., & van Gent, M. R. A. (2021). Individual wave overtopping at coastal structures: A critical review and the existing challenges. *Applied Ocean Research*, *106*(July 2020), 102476. https://doi.org/10.1016/j.apor.2020.102476

Koosheh, A., Etemad-Shahidi, A., Cartwright, N., Tomlinson, R., & van Gent, M. R. A. (2022). Experimental study of wave overtopping at rubble mound seawalls. *Coastal Engineering*, *172*(November 2021), 104062. https://doi.org/10.1016/j.coastaleng.2021.104062

Le, L. T., Nguyen, H., Zhou, J., Dou, J., & Moayedi, H. (2019). Estimating the heating load of buildings for smart city planning using a novel artificial intelligence technique PSO-XGBoost. *Applied Sciences (Switzerland)*, *9*(13). https://doi.org/10.3390/APP9132714

Lim, S., & Chi, S. (2019). Xgboost application on bridge management systems for proactive damage estimation. *Advanced Engineering Informatics*, *41*(October 2018), 100922. https://doi.org/10.1016/j.aei.2019.100922

Longuet-Higgins. (1952). ON THE STATISTICAL DISTRIBUTION OF THE HEIGHTS OF SEA WAVES. *JOURNAL OF MARINE RESEARCH*.

Muslim, T. O., Ahmed, A. N., Malek, M. A., Afan, H. A., Ibrahim, R. K., El-Shafie, A., Sapitang, M., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2020). Investigating the influence of meteorological parameters on the accuracy of sea-level prediction models in Sabah, Malaysia. *Sustainability (Switzerland)*, *12*(3). https://doi.org/10.3390/su12031193

Nayak, P. C., Sudheer, K. P., Rangan, D. M., & Ramasastri, K. S. (2005). Short-term flood forecasting with a neurofuzzy model. *Water Resources Research*, *41*(4), 1–16. https://doi.org/10.1029/2004WR003562

Orme, A. R. (2015). The Four Traditions of Coastal Geomorphology. In *Treatise on Geomorphology* (Vol. 10). https://doi.org/10.1016/B978-0-12-374739-6.00270-0

Owen, M. W. (1981). *Design of seawalls allowing for wave overtopping. June*.

Peregrine, D. H., & Williams, S. M. (2001). Swash overtopping a truncated plane beach. *Journal of Fluid Mechanics*, *440*, 391–399. https://doi.org/10.1017/S002211200100492X

Pillai, K., Etemad-Shahidi, A., & Lemckert, C. (2017). Wave overtopping at berm breakwaters: Review and sensitivity analysis of prediction models. *Coastal Engineering*, *120*(October 2016), 1–21. https://doi.org/10.1016/j.coastaleng.2016.11.003

Pillai, K., Etemad-Shahidi, A., & Lemckert, C. (2019). Wave run-up on bermed coastal structures. *Applied Ocean Research*, *86*(August 2018), 188–194. https://doi.org/10.1016/j.apor.2019.02.006

Pillai, K. K., Etemad-Shahidi, A., & Lemckert, C. (2020). Wave Reflection From Berm Breakwaters. *Coastal Engineering Proceedings*, *36v*, 7. https://doi.org/10.9753/icce.v36v.structures.7

Pullen, T., Allsop, W., Bruce, T., & Geeraerts, J. (2003). Violent wave overtopping: CLASH field measurements at Samphire Hoe. *Coastal Structures 2003 - Proceedings of the Conference*, 469–480. https://doi.org/10.1061/40733(147)39

Rezaie-Balf, M., Naganna, S. R., Kisi, O., & El-Shafie, A. (2019). Enhancing streamflow forecasting using the augmenting ensemble procedure coupled machine learning models: case study of Aswan High Dam. *Hydrological Sciences Journal*, *64*(13), 1629–1646. https://doi.org/10.1080/02626667.2019.1661417

Rogers, R. D. (2020). Machine learning and coastal processes. In *Sandy Beach Morphodynamics* (Issue Chapter 8). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-102927-5/00028-X

Salauddin, M., & Pearson, J. M. (2020). Laboratory investigation of overtopping at a sloping structure with permeable shingle foreshore. *Ocean Engineering*, *197*(November 2019), 106866. https://doi.org/10.1016/j.oceaneng.2019.106866

Shaeri, S., & Etemad-Shahidi, A. (2021). Wave overtopping at vertical and battered smooth impermeable structures. *Coastal Engineering*, *166*(July 2020), 103889. https://doi.org/10.1016/j.coastaleng.2021.103889

Shafaei, M., & Kisi, O. (2016). Lake Level Forecasting Using Wavelet-SVR, Wavelet-ANFIS and Wavelet-ARMA Conjunction Models. *Water Resources Management*, *30*(1), 79–97. https://doi.org/10.1007/s11269-015-1147-z

STEENDAM, G. J., VAN DER MEER, J. W., VERHAEGHE, H., BESLEY, P., FRANCO, L., & VAN GENT, M. R. A. (2005). *the International Database on Wave Overtopping*. *January*, 4301–4313. https://doi.org/10.1142/9789812701916_0347

TAW. (2002). Technical report wave run-up and wave overtopping at dikes. *Technical Advisory Committee on Flood Defence, Delft, The Netherlands*, 43.

Techniek, V. C. (2005). *Voorspelling van golfoverslag over golfbrekers en zeeweringen met behulp van neurale netwerken Neural Network Prediction of Wave Overtopping at Coastal Structures Hadewych Verhaeghe*.

Thorndike Saville, J. (1986). LABORATORY DATA ON WAVE RUN - UP AND OVERTOPPING ON SHORE STRUCTURES by. In *Biologia Centrali-Americaa* (Vol. 2).

van der Meer, J., & Bruce, T. (2014). New Physical Insights and Design Formulas on Wave Overtopping at Sloping and Vertical Structures. *Journal of Waterway, Port, Coastal, and Ocean Engineering*, *140*(6), 1–18. https://doi.org/10.1061/(asce)ww.1943-5460.0000221

van der Meer, J. W. (1998). Application and stability criteria for rock and artificial units. *Dikes and Revetments: Design, Maintenance and Safety Assessment*, *November 2014*, 191–215. https://doi.org/10.1201/9781315141329

van der Meer, J. W. (2017). Geometrical design of coastal structures. *Dikes and Revetments: Design, Maintenance and Safety Assessment*, *January*, 161–175. https://doi.org/10.1201/9781315141329

van der Meer, J. W., Briganti, R., Zanuttigh, B., & Wang, B. (2005). Wave transmission and reflection at low-crested structures: Design formulae, oblique wave attack and spectral change. *Coastal Engineering*, *52*(10–11), 915–929.

https://doi.org/10.1016/j.coastaleng.2005.09.005

van der Meer, J. W., Verhaeghe, H., & Steendam, G. J. (2009). The new wave overtopping database for coastal structures. *Coastal Engineering*, *56*(2), 108–120. https://doi.org/10.1016/j.coastaleng.2008.03.012

van Dongeren, A., Ciavola, P., Martinez, G., Viavattene, C., Bogaard, T., Ferreira, O., Higgins, R., & McCall, R. (2018). Introduction to RISC-KIT: Resilience-increasing strategies for coasts. *Coastal Engineering*, *134*(February 2017), 2–9. https://doi.org/10.1016/j.coastaleng.2017.10.007

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), 586–593. https://doi.org/10.1016/j.coastaleng.2006.12.001

Verhaeghe, H. (2005). *Neural Network Prediction of Wave Overtopping at Coastal Structures Hadewych Verhaeghe*.

Wee, W. J., Zaini, N. B., Ahmed, A. N., & El-Shafie, A. (2021). A review of models for water level forecasting based on machine learning. *Earth Science Informatics*, *14*(4), 1707–1728. https://doi.org/10.1007/s12145-021-00664-9

Williams, H. E., Briganti, R., Romano, A., & Dodd, N. (2019). Experimental analysis of wave overtopping: A new small scale laboratory dataset for the assessment of uncertainty for smooth sloped and vertical coastal structures. *Journal of Marine Science and Engineering*, *7*(7), 1–18. https://doi.org/10.3390/jmse7070217

Zanuttigh, B., Formentin, S. M., & Briganti, R. (2013). A neural network for the prediction of wave reflection from coastal and harbor structures. *Coastal Engineering*, *80*, 49–67. https://doi.org/10.1016/j.coastaleng.2013.05.004

Zanuttigh, B., Formentin, S. M., & van der Meer, J. W. (2016a). Prediction of extreme and tolerable wave overtopping discharges through an advanced neural network. *Ocean Engineering*, *127*(March), 7–22. https://doi.org/10.1016/j.oceaneng.2016.09.032

Zanuttigh, B., Formentin, S. M., & van der Meer, J. W. (2016b). Prediction of extreme and tolerable wave overtopping discharges through an advanced neural network. *Ocean Engineering*, *127*(July), 7–22. https://doi.org/10.1016/j.oceaneng.2016.09.032

Zanuttigh, B., Formentin, S. M., & Van der Meer, J. W. (2017). Update of the Eurotop Neural Network Tool: Improved Prediction of Wave Overtopping. *Coastal Engineering Proceedings*, *35*, 2. https://doi.org/10.9753/icce.v35.waves.2

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost. *IEEE Access*, *6*, 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678

Zhou, J., Qiu, Y., Armaghani, D. J., Zhang, W., Li, C., Zhu, S., & Tarinejad, R. (2021). Predicting TBM penetration rate in hard rock condition: A comparative study among six XGB-based metaheuristic techniques. *Geoscience Frontiers*, *12*(3), 101091. https://doi.org/10.1016/j.gsf.2020.09.020