Journal of Environmental Management 321 (2022) 115923

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Robust machine learning algorithms for predicting coastal water quality index

Md Galal Uddin [a,b,c,*], Stephen Nash [a,b,c], Mir Talas Mahammad Diganta [a,b,c], Azizur Rahman [d,e], Agnieszka I. Olbert [a,b,c]

[a] School of Engineering, National University of Ireland Galway, Ireland
[b] Ryan Institute, National University of Ireland Galway, Ireland
[c] MaREI Research Centre, National University of Ireland Galway, Ireland
[d] School of Computing, Mathematics and Engineering, Charles Sturt University, Wagga Wagga, Australia
[e] The Gulbali Institute of Agriculture, Water and Environment, Charles Sturt University, Wagga Wagga, Australia

ARTICLE INFO

Keywords:
Robust machine learning algorithms
Coastal water quality index model
Coastal water quality
Uncertainty
Cork harbour

ABSTRACT

Coastal water quality assessment is an essential task to keep "good water quality" status for living organisms in coastal ecosystems. The Water quality index (WQI) is a widely used tool to assess water quality but this technique has received much criticism due to the model's reliability and inconsistence. The present study used a recently developed improved WQI model for calculating coastal WQIs in Cork Harbour. The aim of the research is to determine the most reliable and robust machine learning (ML) algorithm(s) to anticipate WQIs at each monitoring point instead of repeatedly employing SI and weight values in order to reduce model uncertainty. In this study, we compared eight commonly used algorithms, including Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGB), Extra Tree (ExT), Support Vector Machine (SVM), Linear Regression (LR), and Gaussian Naïve Bayes (GNB). For the purposes of developing the prediction models, the dataset was divided into two groups: training (70%) and testing (30%), whereas the models were validated using the 10-fold cross-validation method. In order to evaluate the models' performance, the RMSE, MSE, MAE, R^2, and PREI metrics were used in this study. The tree-based DT (RMSE = 0.0, MSE = 0.0, MAE = 0.0, R^2 = 1.0 and PERI = 0.0) and the ExT (RMSE = 0.0, MSE = 0.0, MAE = 0.0, R^2 = 1.0 and PERI = 0.0) and ensemble tree-based XGB (RMSE = 0.0, MSE = 0.0, MAE = 0.0, R^2 = 1.0 and PERI = +0.16 to −0.17) and RF (RMSE = 2.0, MSE = 3.80, MAE = 1.10, R^2 = 0.98, PERI = +3.52 to −25.38) models outperformed other models. The results of model performance and PREI indicate that the DT, ExT, and GXB models could be effective, robust and significantly reduce model uncertainty in predicting WQIs. The findings of this study are also useful for reducing model uncertainty and optimizing the WQM-WQI model architecture for predicting WQI values.

1. Introduction

In any aquatic ecosystem, freshwater is an important bio-indicator for living organisms and therefore, the new challenge for the world's future is to maintain "good water quality status". Recently a few studies have revealed that around 50 marine species including 48 fish species, crustaceans, shellfish, and five types of seaweed living in Irish waters are under threat of extinction due to the water quality and functional changes of habitat of aquatic system (Fogarty P., 2017). Water quality deteriorates over time due to a variety of factors, one of which is human intervention. Industrialization and urbanization have accelerated day by day to ensure a better quality of life. As a consequence, freshwater consumption has significantly increased over many decades (Gikas et al., 2020; Uddin et al., 2018). Therefore, both anthropogenic and natural events have gradually accumulated, resulting in fast degradation of surface and groundwater quality (Aschonitis et al., 2012; Uddin et al., 2020, 2021).

Water resources management is a critical process involving various components, including institutional framework, skilled labour, legislation, financial freedom and resource availability. Several countries have formulated management and action plans to maintain their good water quality. However, due to resource availability, they face a few common

* Corresponding author. Civil Engineering, College of Science and Engineering, National University of Ireland Galway, Ireland.
E-mail address: u.mdgalal1@nuigalway.ie (M.G. Uddin).

https://doi.org/10.1016/j.jenvman.2022.115923
Received 1 June 2022; Received in revised form 6 July 2022; Accepted 30 July 2022
Available online 19 August 2022

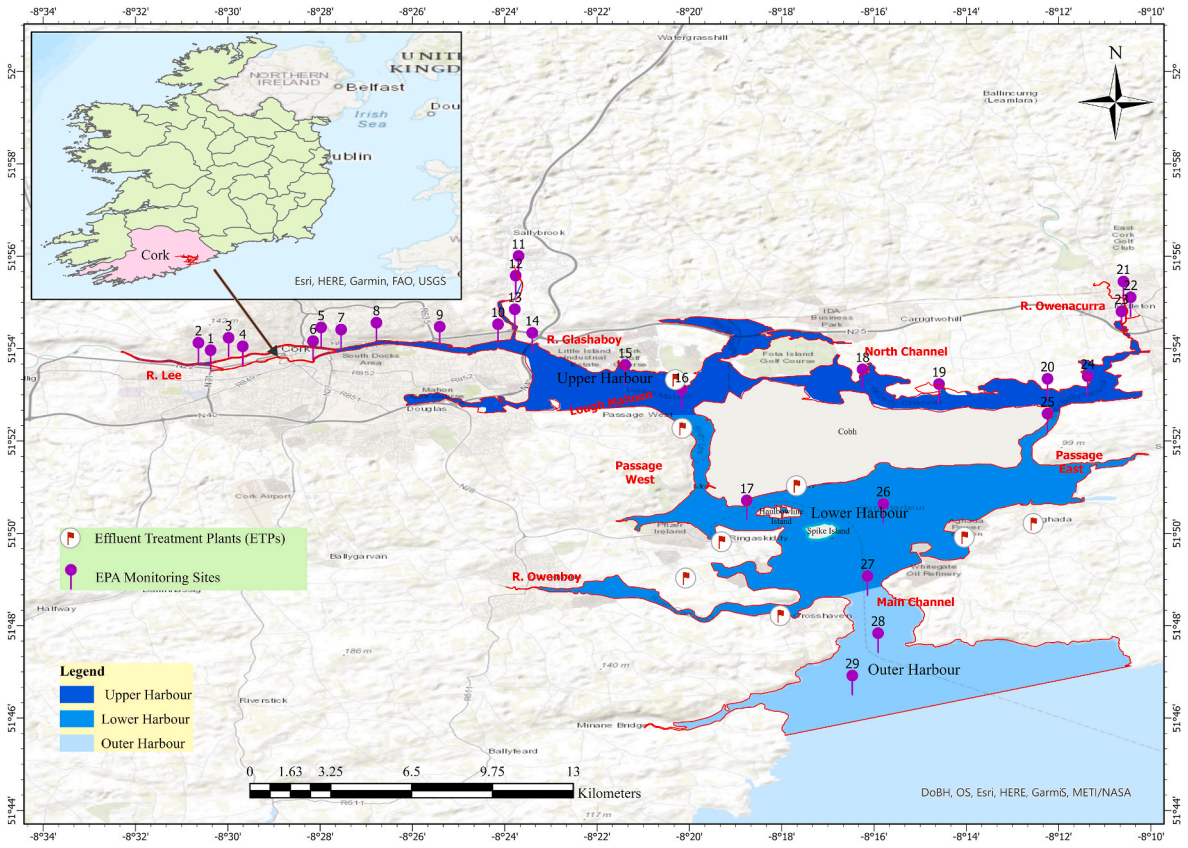

Fig. 1. Study domain: EPA water quality monitoring sites and effluent treatment plants (ETPs) in Cork Harbour, Ireland.

problems in implementing or adopting the management program. In Europe, Water Framework Directive (WFD) is an effective tool for managing water and its ecosystem (Uddin et al., 2022c). It recommended adopting the monitoring program to investigate water quality by all member states as already; many countries have been suffering its challenges and trying to overcome that issue (Zotou et al., 2019).

Thus far, several tools and techniques have been developed for assessing water quality. The water quality index is one of them. Recently, this technique has been extensively used to evaluate water quality. Its application has increased rapidly due to its ability to convert a vast amount of water quality information into a unitless numerical expression using simple mathematical functions. Commonly, this technique consists of four crucial elements: (i) selecting water quality indicator; (ii) sub-index process; (iii) weighting of water quality indicators; and (iv) aggregation function. Further details of the WQI models and their uses are available in the literature (e.g., Uddin et al., 2021). Recently, several studies have revealed that the WQI model produced considerable uncertainty in its modeling process (Abbasi and Abbasi, 2012; Juwana et al., 2016; Rezaie-Balf et al., 2020; Sutadian et al., 2016; Uddin et al., 2021). As a result, the WQI model does not reflect accurate water quality attributes. Many researchers have proposed a range of modified WQIs for optimizing this issue, but unfortunately, several recent studies have revealed that those models have experienced similar problems (Abbasi and Abbasi, 2011; Chang et al., 2020; Smith, 1990; Stoner, 1978; Yan et al., 2016).

Also, this method is much more sensitive to eclipsing and ambiguity problems. The "eclipsing" problem can be occurred due to inappropriate sub-indexing rules, parameter weightings, or inappropriate aggregation functions that do not reflect the real information of water quality (Sutadian et al., 2016; Uddin et al., 2021; 2022c). Recently, a few studies have revealed that the "eclipsing" problem occurs due to overestimation of the WQI index by the aggregation function (Uddin et al., 2022c. Like eclipsing, ambiguity is another important source of the WQI model uncertainty. It hides the actual water quality information by underestimation and overestimation of WQI values (Uddin et al., 2021, Uddin et al., 2022c). Details of the eclipsing and ambiguity problems discussed by Uddin et al., 2022c. In the WQI model, several studies have considered the effects of ambiguity and eclipsing issues of sub-index and aggregation functions (Smith, 1990; Abbasi and Abbasi, 2012). Details of the ambiguity and eclipsing problems, sources and impact on WQI model of them can be found in Uddin et al., 2022c. To determine their effects, Uddin et al. (2022) compared eight WQI models (four weighted and four unweighted) to evaluate the ambiguity and eclipsing problems for assessing coastal water quality in his study. This study recommended that the WQM-WQI model could be effective and reliable for assessing coastal water quality in terms of reducing uncertainty in the WQI model.

Due to the inconsistency of existing WQI techniques, a few researchers have recently used the ML technique to reduce model uncertainty and attempt to predict WQIs accurately (Babbar and Babbar, 2017; Bui et al., 2020; Gao et al., 2020; Hassan et al., 2021; Kouadri et al., 2021; Rezaie-Balf et al., 2020; Wang et al., 2017). Several studies have applied a variety of ML algorithms such as extreme gradient boosting, Naïve Bayes, support vector machine, random forest, and decision tree algorithms for the comparison of algorithms performance in order to predict WQIs correctly (Ahmad et al., 2017; Bui et al., 2020; Deng et al., 2022; Khan and See, 2016; Leong et al., 2019; Othman et al., 2020). A summary of the various ML techniques in predicting water quality is provided in Annex (a). Bui et al. (2020) compare sixteen algorithms to identify the robust model for predicting WQIs accurately. They suggest that tree-based algorithms are practical for predicting WQIs. Some studies recommend that ensemble tree-based algorithms such as extreme gradient boosting (XGB) and random forest (RF) are potentially useful for predicting WQIs (Grbčić et al., 2021; Haghiabi et al., 2018a; Islam Khan et al., 2021; Khullar and Singh, 2021). Moreover, researchers successfully applied AI-based algorithms like support vector machine (SVM), least square SVM (LSVM) and artificial neural

Table 1
Water quality parameters, units and standard threshold for coastal water quality accordance to Uddin et al., 2022c.

Parameter	Unit	Standard threshold	
		Lower	Upper
CHL[a]	mg/m^3	0.0	14.2
DOX[a]	% sat	72	128
MRP[a]	mg/l as P	0.0	0.05
DIN[a]	mg/l	0.0	1.20
AMN[b]	mg/l	0	1.5
BOD[b]	mg/l	0	7
pH[c]	–	5	9
TEMP[b]	°C	–	25
TON[d]	mg/l as N	0.0	2
TRAN[e]	m/depth	>1	–
SAL[b]	psu	12	38

[a] ATSEBI guide values, indicators standard values was obtained based on median value of salinity. In this study, SAL median value was found 20.47 (see details in Annex 1 d).
[b] EPA, Ireland (2001), recommended values for the surface water.
[c] pH and Alkalinity Monitoring Manual for estuary, EPA,USA.
[d] The European Communities regulations for quality of surface water intended for the abstraction of drinking water, 1989 (S.I. No. 294/1989).
[e] EPA′ bathing Water Quality Regulations 2008, (Ref. No. 79/2008).

network for predicting WQIs (Aldhyani et al., 2020; Haghiabi et al., 2018b; Pham et al., 2019; Prasad et al., 2022; Wu and Wang, 2022). However, most studies have focused on the river/lake or groundwater quality index utilizing the existing WQI models while most studies have been carried out on only the prediction of WQIs, no studies have been found to improve the WQI model architecture. Compared to other studies, the present research widely explores, for the first time, to improve the newly developed WQM-WQI model architecture using ML techniques in order to reduce the model uncertainity.

This study aims to identify the robust ML algorithm with optimizing the hyperparameters for predicting WQIs correctly at each monitoring site in Cork Harbour, Ireland, comparing eight widely used ML algorithms Decision Tree (DT), Extra Tree (ExT), Extreme Gradient Boosting (XGB), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Linear Regression (LR), and Gaussian Naïve Bayes (GNB). We use these algorithms to determine outperformed models to reduce the WQI model prediction uncertainty and improve the model architecture especially coastal WQIs.

The paper is developed as follows: Section 2 presents the details of the nature of Cork Harbour and its environmental significance. Section 3 describes the details overview of various ML algorithms, validation processes and other statistical methods for assessing model, Section 4 provides in depth of the prediction results and discusses the output of the prediction models, and Section 5 summarizes the findings, recommendations, limitations and future direction of this research.

2. Application domain- a case study in Cork Harbour

The present study was conducted in Cork Harbour as Special Protection Area (SPA), that is relatively the deepest and longest (17.72 km) surface waterbodies in Ireland (Hartnett and Nash, 2015; Nash et al., 2011). The Harbour has covered with large surface area (85.85 km^2) and brackish estuary on the south coast of Ireland (Nash et al., 2011). It is a macro-tidal with a typical spring tide range of 4.2 m at the entrance to the Harbour (Uddin et al., 2022c). Relatively, the Cork city is the well-known as an industrial hub of Ireland and the surrounding hinterlands are dominated with extensive agricultural practices which influence water quality in the region directly due to the using chemical fertilizers for developing crops (EPA, 2017). Recently, several annual environmental reports of EPA has revealed that the Cork and Donegal received the highest raw discharge waste water directly without any treatment (EPA, 2017). Moreover, the Cork Harbors' geological patterns

Table 2
Classification scheme for coastal water quality.

Classification scheme	Range of score	Description
Good	80–100	Water quality is suitable to use for any purposes.
Fair	50–79	A few indicators meet the guide values and the water quality is safe with a minor observation.
Marginal	30–49	Most of the indicators does not fall into the criteria; water quality is unsafe, which may be harmful for aquatic life.
Poor	0–29	Each indicators failed to meet all the criteria; water quality is completely unsafe and not suitable for many certain uses.

are vital for Harbour area's ecosystem and fresh water quality. It has been identified as a Special Protection Area (SPA) under the 1979 Wild birds Directive (79/409/EEC).

3. Methods and materials

3.1. Data obtaining process

Water quality data was retrieved from the Irish Environmental Protection Agency (EPA) water quality monitoring database for Cork Harbour. The details of the data are available at https://www.catchments.ie/data. Typically, the EPA monitors the water quality of Habour frequently. A total of 29 monitoring locations out of 32 were considered for this study. Details of the monitoring sites and their descriptions are provided in Annex 1(b). Fig. 1 provides the details of the monitoring locations in Cork Harbour. This study uses eleven water quality variables for the WQI calculation: temperature (TEMP), total organic nitrogen (TON), ammonia (AMN), dissolved oxygen (DOX), ammoniacal nitrogen (AMN), pH, salinity (SAL), molybdate reactive phosphorus (MRP), biological oxygen demand (BOD), transparency (TRAN), and *Chlorophyll a* (CHL). Selected WQ indicators data was considered from 1 m depth at each monitoring site in Cork Harbour. Table 1 provides an overview of the studied water quality indicators unit; standard threshold and Annex 1(c) supply the details of the indicator monitoring data at each site, respectively. Water quality indicators were considered for this study based on the availability of data variables in the monitoring database 2020, considering the fine dissemination of monitoring sites. For further analysis, averaged concentrations (from January 2020 to December 2020) of indicators were used in this research for further analysis (Annex 1c).

3.2. WQI calculation

A range of WQI models has been used to calculate the WQI values. Its application has increased sequentially due to its simple mathematical functions and ease of use. However, the existing literature on WQI models is extensive and focuses mainly on details of WQI models and their services (Gupta and Gupta, 2021; Uddin et al., 2021) without checking their statistical accuracy. Hence, this research enhances the accuracy of the WQM based WQI model outputs by estimating more precise and statistically reliable water quality index scores. Typically, an ideal WQI model comprises four components such as water indicators selection, sub-index (SI) function, indicators weight generation and aggregation function. Details of these components with statistical functions are available in the literature (e.g., Rahman and Harding, 2016; Uddin et al., 2022c; Uddin et al., 2022a, 2022b). In this study, the WQIs was calculated based on the improvement methodology proposed by Uddin et al., 2022c because this approach is one of the more practical and effective for assessing coastal water quality. In details, the weighted quadratic mean (WQM) WQI methodology can be found in Uddin et al., 2022c.

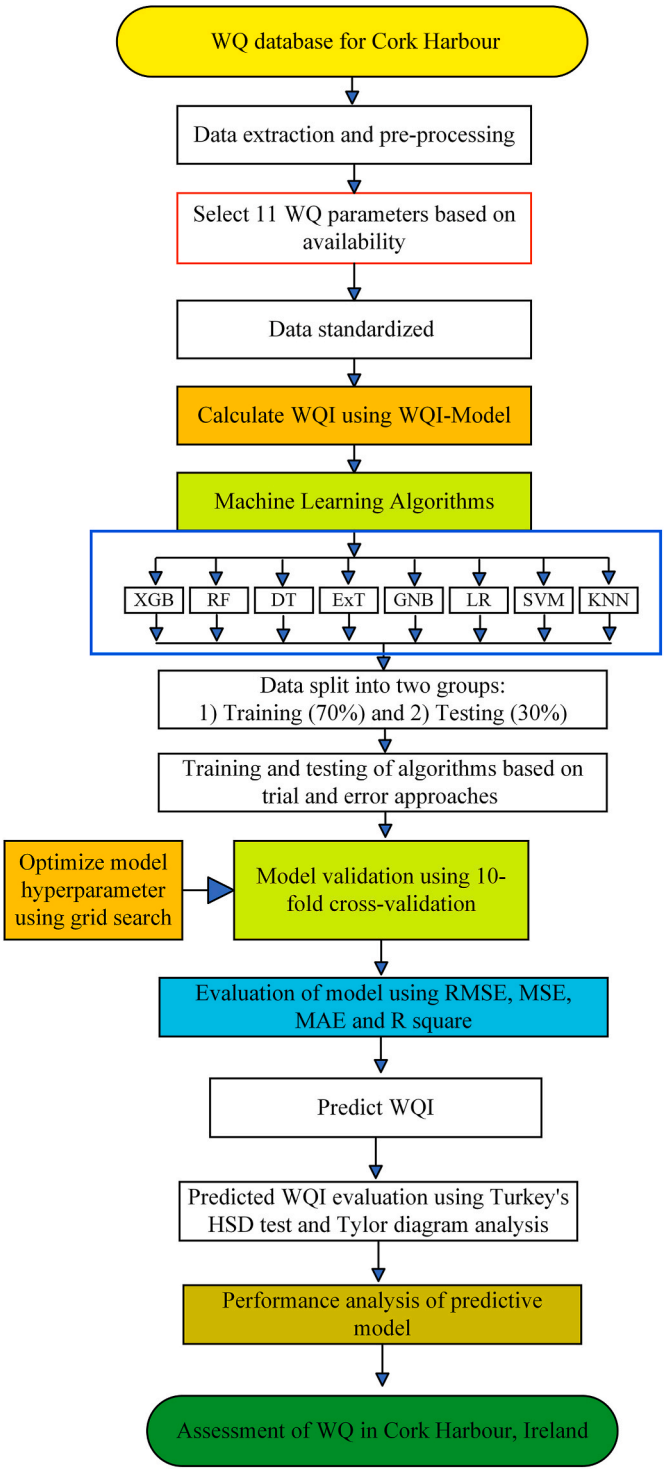

Fig. 2. A comprehensive framework for the assessment of predicting WQIs.

3.2.1. Evaluation of WQIs

Currently, various classification schemes are used to evaluate the WQIs in literature. Several recent studies have claimed that the WQIs model results do not reflect the actual information on water quality due to the various classification schemes for similar data attributes (e.g., see Uddin et al., 2022c). Uddin et al., 2022c proposed unique classification schemes for assessing coastal water quality, and we used these schemes in this research. Table 2 provides the details of the classification schemes.

3.3. Data pre-processing

3.3.1. Data standardization

Prior to the training of ML algorithms, it is essential to standardize data variables. Commonly, in ML technique, this method used for converting all data variables into a uniform scale in order to optimise the model training errors (Rahman, 2019, 2020; Solanki et al., 2015). In this study, water quality data variables were standardized using z score normalization process. Z score can be presented as follows:

$$z = \frac{x_i - \overline{x}}{\sigma} \quad (1)$$

where, z is the standardize score, x_i is the ith data variable, $\overline{x}$ refers to the mean of data variable and σ is the standard deviation of data.

3.3.2. Data splitting

Before, training the ML algorithms, data was divided into training [70% (20 monitoring sites)] and testing [30% (9 monitoring sites)] sets. After splitting data, eight ML algorithms were trained and tested using training and testing data sets respectively. Model performance was evaluated for both phases.

3.4. Machine learning algorithms

ML technique is widely used to predict unknown objects. Recently, this technique has been utilized in different branches of research. For example, research on predicting water quality has revealed that the ML algorithm could be more effective in evaluating water quality than other traditional methods (Aldhyani et al., 2020; Azrour et al., 2021; Babbar and Babbar, 2017; Haghiabi et al., 2018a; Mohammed et al., 2018; Prakash et al., 2018; Solanki et al., 2015; Xiong et al., 2020). Several studies have effectively used machine learning approaches to predict WQI (Ahmad et al., 2017; Bui et al., 2020; Grbčić et al., 2021; Hassan et al., 2021; Kadam et al., 2019; Kouadri et al., 2021; Leong et al., 2019; Venkata Vara Prasad et al., 2020; Wang et al., 2017). This research utilized eight ML algorithms to identify robust algorithms for predicting WQM-WQIs. The details methodological procedures of this study are presented in Fig. 2. The details of various ML algorithms can be found in the supplementary material as a continuation of 3.4.1.

3.4.1. Model hyper-parameterization

Hyper-parameters tuning of ML technique is performed to obtain higher level model accuracy (Elgeldawi et al., 2021; Villalobos-Arias et al., 2020). In ML approaches, numerous methods are used to hyper-parameterise the predictive model. Most of the studies in the literature used grid search and random search techniques to optimise the model hyper-parameters (Shekar and Dagnew, 2019). The grid search technique is widely used because this technique evaluates model accuracy for each grid position (Elgeldawi et al., 2021; Shekar and Dagnew, 2019). Compared to the typical hyper-parametrization process, the grid search is more efficient method than the random search (Villalobos-Arias et al., 2020). Thus, this research uses the grid search technique to optimise the model parameters. Table 3 presents hyper-parameters for various ML models the during model training phase.

3.4.2. Model performance analysis

3.4.2.1. Cross-validation approaches. Cross-validation (CV) is the most common procedures to evaluate the ML models for small datasets. To assess the performance of ML predictive model, the present study is used the random CV technique to compare the model performance. In this study, 10-fold CV technique was utilized including widely used four evaluation criteria: mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Details technique can be found in Xiong et al. (2020). Except for R^2,

Table 3
Optimized hyper-parameters of various ML models during testing period.

Model parameters	XGB	RF	DT	ExT	LR	KNN	SVM	GNB
n_estimators	100	100	100	100	100	–	30	200
learning_rate	0.2	–	–	–	–	–	–	–
max.depth	20	10	10	20				–
gamma	0	–	–	–	–	–	auto	–
booster	gbtree	–	–	–	–	–		–
Kernel		–	–	–	–	–	RBF	–
subsample	1	–	–	–	–	–	–	–
colsample_bytree	1	–	–	–	–	–	–	–
base_score	0.5	–	–	–	–	–	–	–
reg_lamda	1	–	–	–	–	–	–	–
bootstrap	True	True	–	True	–	–	–	–
cv.folds	10		–	–	–	–	–	–
random_state	1	1	–	1	–	–	–	–
Objective	reg.linear		–	–	–	–	–	–
criterion	–	Squared_error	–	Squared_error	–	–	–	–
max_leaf_nodes		5	10	5	–	30	–	–
min_samples_leaf		1	5	1	–	–	–	–
epsilon	–	–	–	–	–	–	0.1	–
shrinking	–	–	–	–	–	–	True	–
fit_intercept	–	–	–	–	TRUE	–	–	–
n_neighbors	–	–	–	–	–	5	–	–
weight	–	–	–	–	–	uniform	–	–
metrics	–	–	–	–	–	minkowski	–	–
power_parameters	–	–	–	–	–	2	–	–

the performance criteria expect a predictive model's performance to be as small as possible. In general, the R^2 value refers to assessing the models how well fitted the model with predicted data. It should be close to 1 (He et al., 2015; Sharif et al., 2022). Model evaluation criteria are measured as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

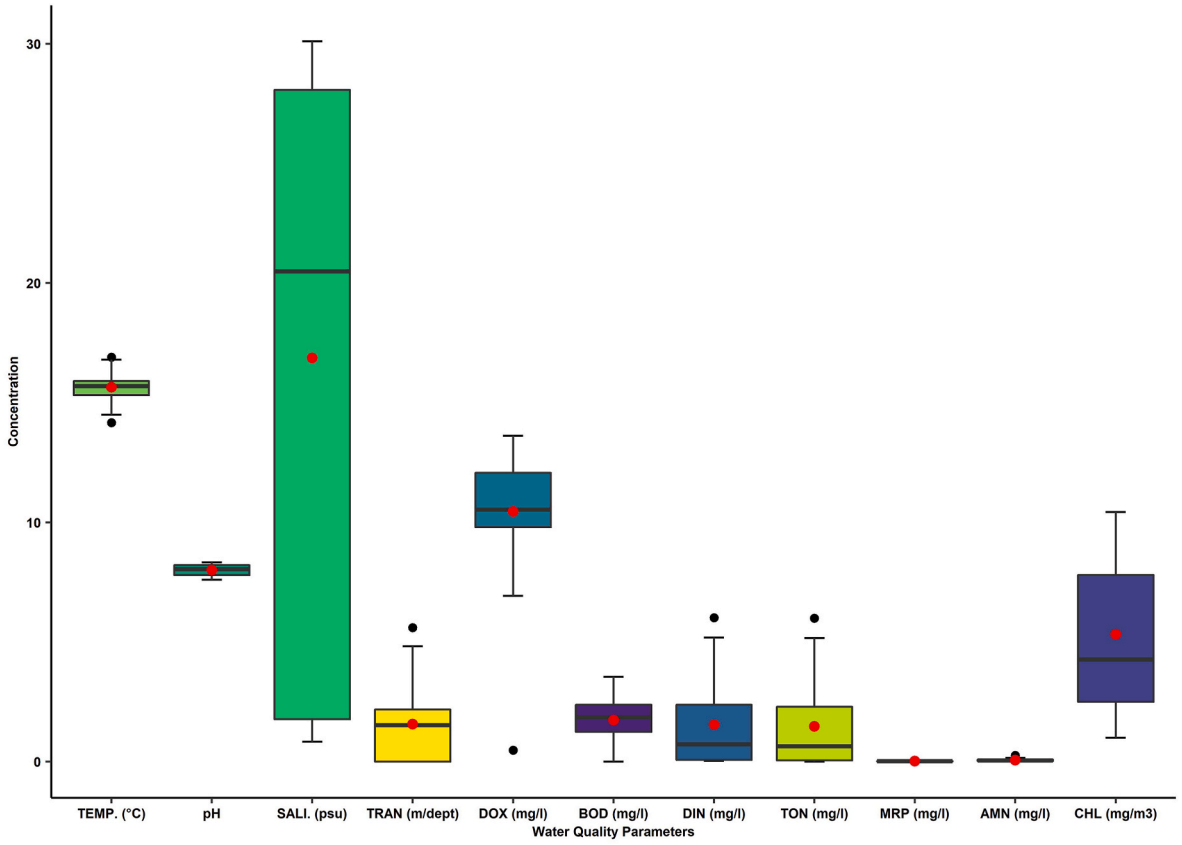

Fig. 3. Physico-chemical attributes of water quality in Cork Harbour.

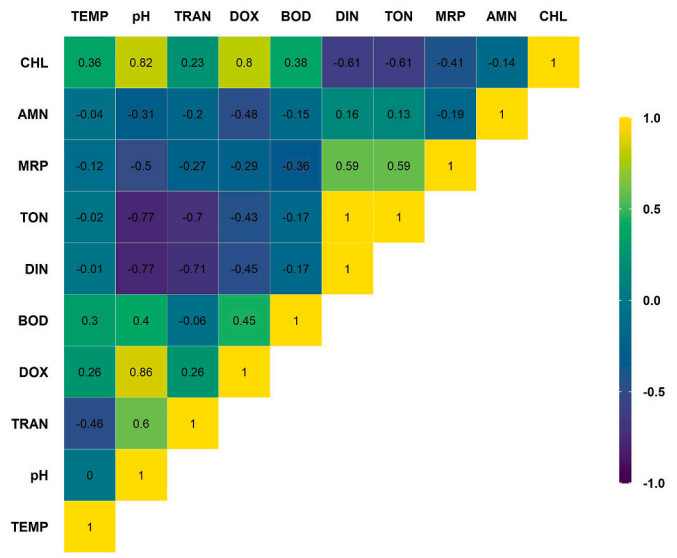

Fig. 4. Pearson's correlation of physico-chemical indicators in Cork Harbour.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{10}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \widehat{y})^2} \tag{11}$$

where y_i and $\widehat{y}$ are the *i*th observed and mean of the predicted values respectively. *N* is the number of observations.

3.4.2.2. Prediction uncertainty analysis. For the purposes of uncertainty analysis in the predictive WQIs of various ML models, several techniques are used, such as Monte Carlo simulation, ML algorithms, etc. In this study, the percent of relative error index (PREI) was utilized to evaluate the prediction error at each observation location because this technique has recently been used for the assessment of predictive bias in predicting ML models (Bui et al., 2020). The result is given in percentage (%). The optimal value of PBIAS is 0.0, with low-magnitude values refers accurate model simulation. Positive values indicate underestimated bias, whereas negative values represent model overestimation bias. Fig. 11 presents the prediction percentage of bias and percent of relative error index, respectively. The PREI are can be defined as follows:

$$\boldsymbol{PREI} = \left(\frac{\boldsymbol{y_i} - \widehat{\boldsymbol{y_i}}}{\boldsymbol{y_i}}\right) \times 100 \tag{12}$$

where y_i actual WQIs for *i*th observation and $\widehat{y_i}$ is the mean predicted WQIs.

In addition, the present study utilized the inferential error bars analysis technique because many studies have utilized this method to evaluate the uncertainty of various datasets or groups. The details of the methodology can be found in Cumming et al. (2007). Fig. 11 presents the uncertainty results of the WQI scores obtained from the various prediction models.

3.4.2.3. Comparative analysis of predictive models. In this study, predictive model bias was analysed by comparing eight ML models using the Tylor diagram. This technique is commonly used to compare various methods, techniques, or models in terms of data deviation. It is effective to identify an appropriate model because it allows three statistical measures, including the correlation between observations and predictions, the root-mean-square deviation (RMSD) and their standard deviations (SD) which help in understanding the model reliability (Calim et al., 2018). Recently, several studies have applied this method to compare the bias among models (Seifi et al., 2020; Xu et al., 2016). Fig. 14 presents the summary of statistics for various ML predictive models.

4. Results

4.1. Physico-chemical assessment of water quality

Fig. 3 presents the descriptive statistics for the studied 11 physico-chemical water quality indicators in Cork Harbour. Basic statistics were obtained using Whisker's box-plot technique, where a black solid line and a red point indicated the median and mean values of water quality indicators, respectively. For the determination of correlation among water quality indicators, the significant associations among the water quality indicators were analysed through the Pearson's correlation test at a 99% confidence level, and the results of the correlation between indicators are presented in Fig. 4.

In this study, water TEMP, pH, CHL, and AMN were found within the standard threshold values of coastal water quality. The details of the standard thresholds are provided in Table 1 above. The highest water TEMP was found at 16.90 °C and the lowest at 13.90 °C with a mean and median value of 15.58 °C and 15.7 °C, respectively throughout the study period, implying a negative skew (mean < median) within the data (Fig. 3). Similarly, water pH also showed a negative skew, having a mean value of 8.00 and a median value of 8.05 (Fig. 3). The CHL ranged from 1.00 mg/m^3 to 10.43 mg/m^3 with a mean value of 5.32 mg/m^3, whereas the AMN's mean concentration was found to be 0.07 mg/l across the monitoring sites in Cork harbour. The values of TRAN were ranged from 0.00 m/depth to 5.60 m/depth with a positive skew within the dataset (mean > median) and showed a significant moderate positive relationship with water pH ($r = 0.60$, $p < 0.01$) (Fig. 4). A significant variation of SAL concentration was observed across the monitoring sites in this research. It varied from 0.83 to 30.1 psμ with a mean value of 16.87 psμ (Fig. 3). It is noted that SAL concentration is only used to determine the standard threshold of coastal water quality for MRP, DOX, CHL and DIN. The details of the procedures can be found in Uddin et al., 2022c. Excessive dissolved oxygen concertation is harmful for aquatic species (Chiang et al., 2021). DOX concentration was found in the North-channel of the upper Harbour, it was 13.61 mg/l (exceed), whereas the lowest concentration (6.93 mg/l) was found in River Lee of the upper Harbour (Annex 1(c)). The DOX showed a significant, strong positive association with water pH ($r = 0.86$, $p < 0.01$) (Fig. 4). The data for BOD showed a higher median value (1.86 mg/l) than the mean value (1.74 mg/l), and it varied between 0.00 mg/l and 3.55 mg/l (Fig. 3).

More than 40% of monitoring sites' DIN concentrations exceeded the upper threshold limit of 1.20 mg/l, with a mean value of 1.54 mg/l (Fig. 3). TON also showed positive skewness across the monitoring sites in Harbour while around 34% of the data points exceeded the upper guideline value of 2 mg/l (Fig. 3). As shown in Fig. 4, a significant strong negative association was found for both DIN and TON with water pH and TRAN, respectively ($r = -0.77$, $p < 0.01$) and ($r = -0.70$, $p < 0.01$). Likely, MRP concentration exceeded the standard threshold of coastal water quality; it ranged from 0.01 mg/l to 0.06 mg/l. The MRP showed a moderate positive relationship with both DIN and TON ($r = 0.59$, p 0.01) (Fig. 4). In this research, a positive correlation was observed between CHL and water pH ($r = 0.82$, $p < 0.01$) and DOX ($r = 0.80$, $p < 0.01$), and a negative relationship was associated with DIN and TON ($r = 0.61$, $p < 0.01$) (Fig. 4).

4.2. Assessing water quality using WQM-WQI models

Water resource management is critical for all states to maintain "good water quality" status. Now a day, the WQI model is widely used to evaluate water quality due to its simple application and easy to evaluate the outcomes of the model. Annex 1(c) provide the details WQIs for each monitoring sites in Cork Harbour. For the evaluation of coastal water

Table 4
Point evaluation of water quality in Cork Harbour using WQM model.

Model	Total monitoring locations	Water quality status			
WQM-WQI	29	Good 0	Fair 44.82% (13)	Marginal 55.18% (16)	Poor 0

quality, the present study was utilized the WQM-WQI model to calculate the WQIs values. It ranged from 33 to 73, with an average of 56.19. Water quality status was evaluated using the coastal water quality classification scheme that are provided in Table 2 above. Water quality status are provided in Table 4 below. Two types of water quality were found in Cork Harbour. These varied from "marginal" to "fair" categories. From Table 4, it can be obtained that in total, 13 (44.82%) of monitoring sites' water quality was found to be "fair", whereas 16 (55.18%) were assessed as "marginal" in Cork Harbour, respectively.

Fig. 5 presents water quality status at each monitoring location in Cork Harbour over the study period. As can be seen from figure below, the Harbour water quality was dominated by the "fair" category, most of the monitoring locations water quality were evaluated as fair quality. The "marginal" class water quality was evaluated in the upper Lee estuary and the upper part of the river Owenacurra (Midleton). The results of the WQM-WQIs also in line with our earlier observations, which showed that the upper Harbour water quality was worst compared to the other parts of the Harbour (Uddin et al., 2022c).

4.3. Comparative analysis of various ML regression models

In this study, we applied eight ML regression algorithms to predict the WQM-WQIs values in Cork Harbour. Annex 1(e) provides the predicted WQM-WQIs for various ML models. In order to validate the predictive results of various ML algorithms, the CV approaches was utilized. Fig. 6 presents the CV results (RMSE, MSE and MAE) for the eight ML models. According to the cross-validation results, the XGB, RF, DT and ExT have the highest prediction perform among the algorithms. The lowest prediction errors belonged to the XGB, DT, and ExT algorithms, but during the training and testing periods the lowest errors were found for the XGB model, whereas the lowest training (RMSE = 3.3, MSE = 10.91, and MAE = 1.67) and testing (RMSE = 0.0, MSE = 0.0, and MAE = 0.02) errors were found for the XGB model. Compared to best algorithms, relatively, higher prediction errors were found for the DT (RMSE = 3.97, MSE = 15.82 and MAE = 2.62) and the ExT (RMSE = 3.60, MSE = 12.96 and MAE = 2.29), respectively during model training period. Interestingly, there was no prediction errors (RMSE, MSE and MAE were found 0 respectively) during testing period, whereas both algorithms were predicted WQI values at each monitoring sites properly. Similar finding also revealed by Chicco et al. (2021). Contrary, the RF also had a very low prediction error, while the GNB and KNN algorithms had higher prediction errors, it ranged from 0.02 to ±3.75. These algorithms did not predict each WQI value accurately at each monitoring site in Harbour. Compared to other algorithms, the SVM performance had very poor over the study period. The large error was found for the SVM model, whereas testing (RMSE, MSE and MAE were 13.40, 179.61 and 12.77 respectively) and training errors also (RMSE, MSE and MAE were 12.68, 160.93 and 11.59 respectively) had higher than other algorithms (Fig. 6).

For the identification of the best algorithms, the present research was also utilized the determination of coefficient (R^2) to evaluate the model performance. Usually, the R^2 refers to the correlation and performance

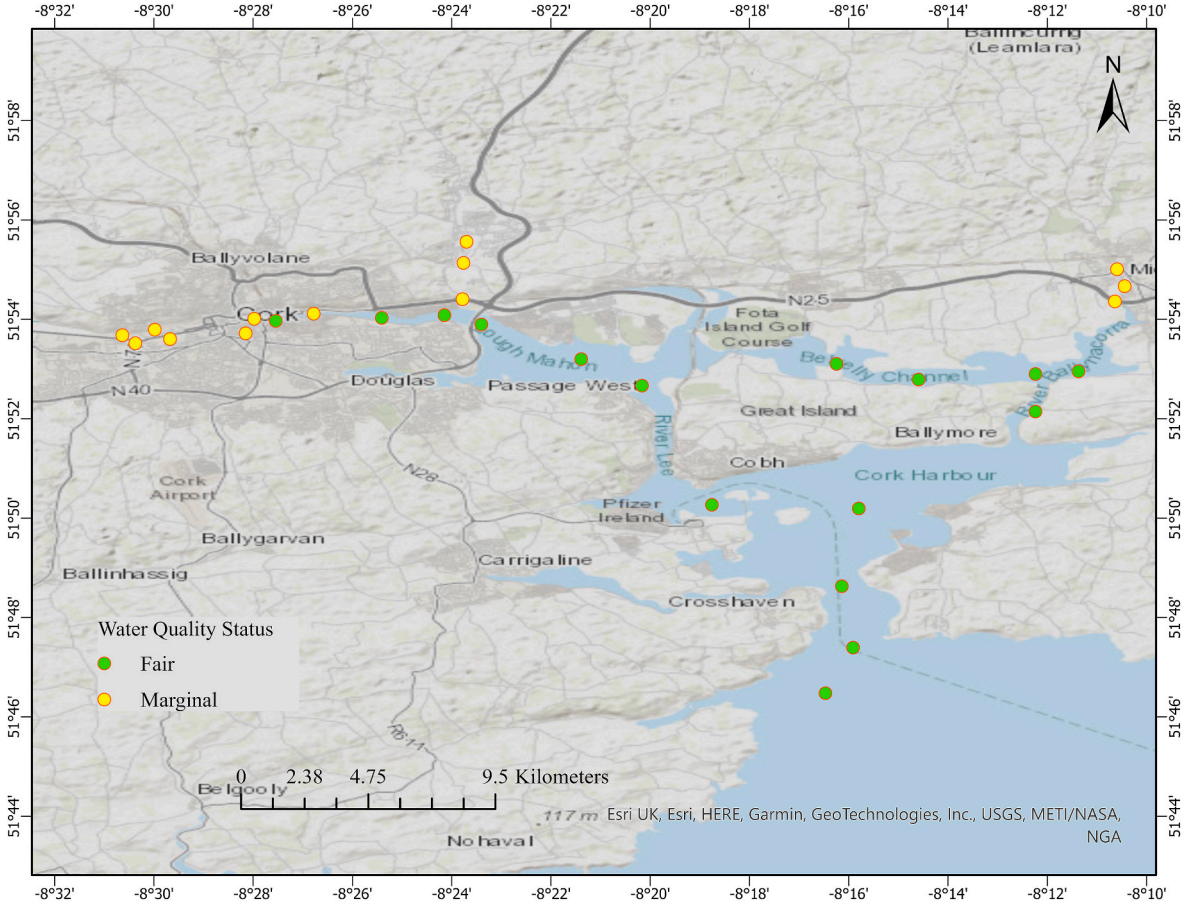

Fig. 5. Water quality status in Cork Harbour using WQM-WQI model.

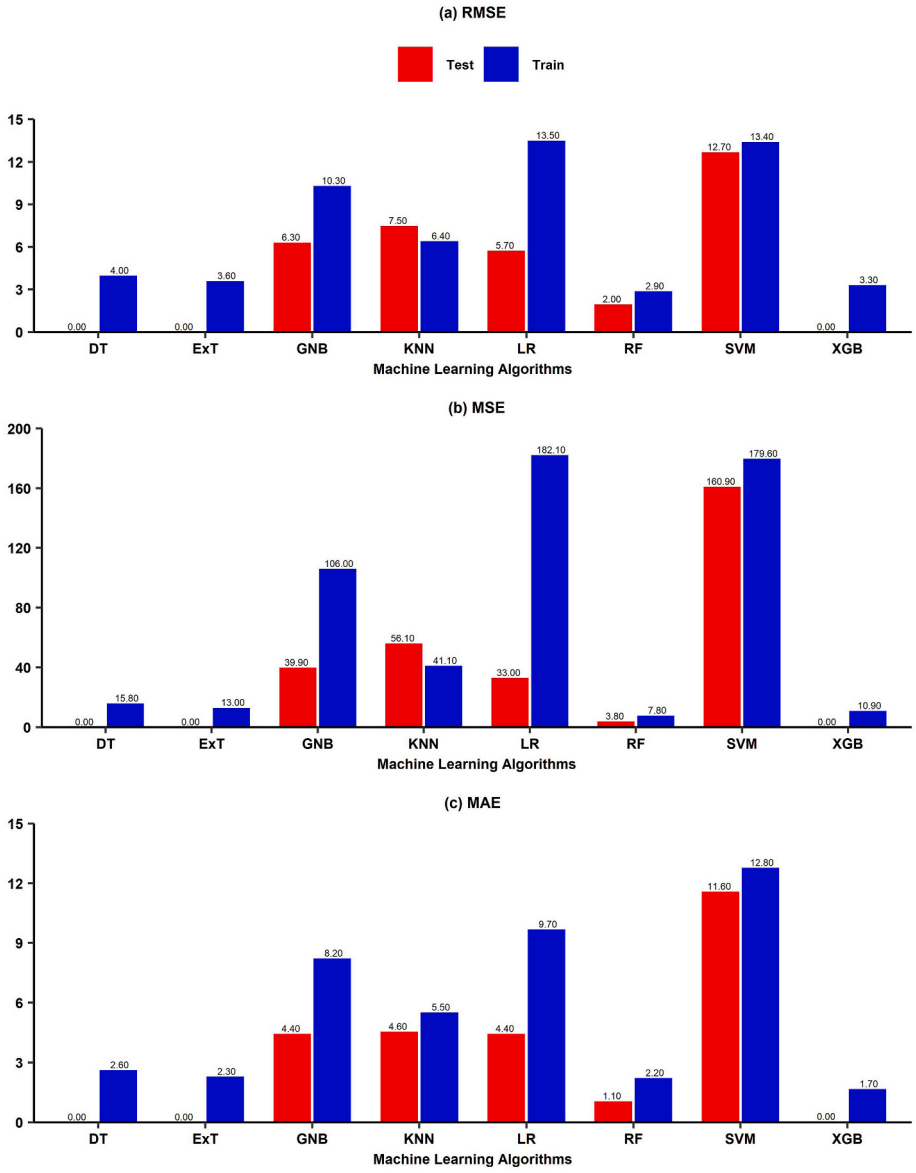

Fig. 6. 10-fold cross-validation results of various ML algorithms.

reliability between predictors and response variables, which helps to identify the best algorithm.

Fig. 7 presents the R^2 value for the various ML algorithms. As discussed above, the XGB, DT and ExT algorithms had lower prediction errors. Those algorithms predicted WQI values at each monitoring sites properly. For these algorithms, R^2 values were found 1, where relative close to 1 (0.98) found for the RF (Fig. 7d). It can be seen from Fig. 7g, there were no relationship between predictors and response for the SVM model whereas the R^2 value had less than 50% (0.43). Although, the LR algorithm had higher R^2 values (0.81) but this algorithm did not predict WQI values at each point properly (Fig. 7f). On the other hand, the KNN and the GNB showed moderate relationship between actual and predicted WQI values (Fig. 7b; Fig. 7h). Therefore, results of R^2 also indicates that the XGB, DT and ExT algorithms had higher predictive capabilities for predicting WQI values.

However, based on the prediction errors, the XGB, DT and ExT were predicted at each WQI values properly. Fig. 8 shows a comparison scenario between actual and predicted WQIs at each monitoring site in Cork Harbour. As can be seen from Fig. 8, all algorithms performed well, except the SVM. Unlike, it showed the worst performed and did not follow the trend to the actual WQI values at each monitoring sites.

In addition, an overview of statistical summary for the predicted WQIs and actual WQIs for various ML models are presented in Fig. 9. Whereas boxplots show differences in predicted WQI values among ML models in line with the actual WQI (Fig. 9a). Fig. 8 also reveals that there has been a slight statistical variation among ML algorithms predicted WQIs except XGB, DT and ExT. As shown in Fig. 9a, compared to all prediction models, there were no significant statistical variation between actual and predictive WQI values for the XGB, DT and ExT models at $p < 0.05$. Completely, different trend was found for the SVM predicted and actual WQI values over the study period, whereas a slight variation was found for the KNN and LR models (Fig. 9a). The Cumulative Distribution Function (CDF) results of the predicted ML models are shown in Fig. 9b. The CDF results indicated that 95% of monitoring sites were predicted correctly except the SVM methods.

Here, we compared several ML models using Tukey's HSD comparison technique. Fig. 10 presents the overall and pair-wise comparison among ML models with a 95% confidence level. The results of Tukey's reveal that there were no statistically significant difference among models. Moreover, the 95% individual confidence level also indicates that predicted WQI values of all models were found between −10 and +10 that means all pairs of predicted means included zero, which

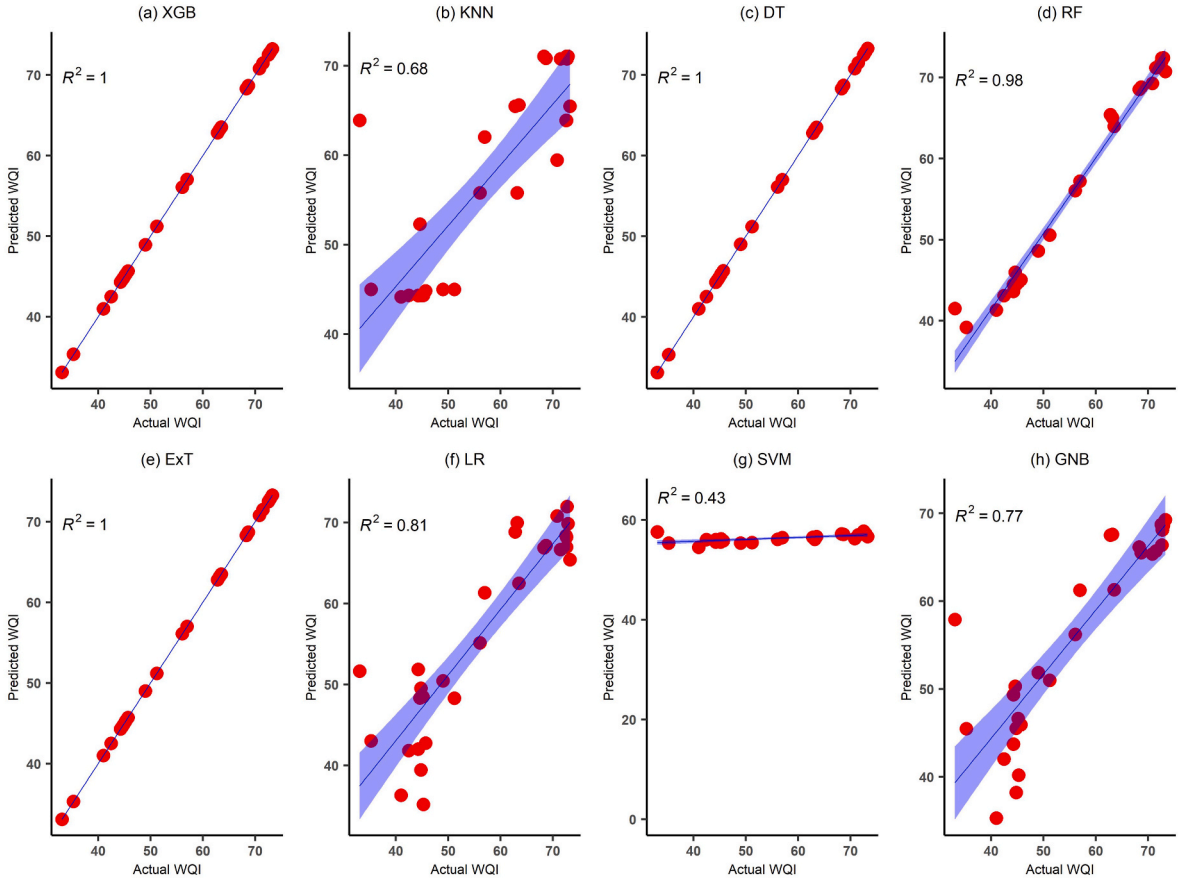

Fig. 7. Scatter plots of actual vs predicted WQI values based on the model testing dataset of different ML regression algorithms for validation purposes.

indicates that the differences are not statistically significant.

4.4. Assessment of uncertainty in predicting WQIs

Based on the CV and HSD analysis, it is hard to determine the best algorithm for predicting WQIs. Here, we utilized PREI, whereas PREI allows comparing the prediction capabilities of models based on their tendency to over or underestimate the WQIs at each observation. It offers an opportunity to examine the prediction power of the model at each data point. Fig. 11 presents the PREI of various predictive models at each location in Cork Harbour.

Compared to among ML algorithms, the lowest error was found for the DT, ExT, XGB, and RF models, respectively. As shown in Fig. 11, underestimating and overestimating biases highly influenced the GNB (+14.71 to −74.9), KNN (+16 to −93), LR (+22.30 to −56.1), RF (+3.52 to −25.38), and SVM (+22.7 to −75) models, respectively. The lowest underestimate and overestimate were found for the XGB model. It ranged from +0.16 to −0.17; whereas there was no bias for DT and ExT in predicting WQIs at each monitoring site. Except XGB, most of the algorithms had overestimated problems in predicting WQIs at monitoring sites in the upper-Eastern part of the Harbour.

However, this study also utilized the 95% confidence interval analysis of predicted WQIs for various ML models. Above, Fig. 12 reveals that similar data discrepancies had the DT, ExT, and XGB models in predicting WQIs. It can be easily figured out from the figure above that there was no data variation between actual (green error bar) and predicted WQIs for these models. In contrast, the GBN, KNN, and LR showed similar errors in predicting WQIs. Unlike GBN, KNN, and LR, comparatively low errors were made in predicting WQIs for the RF model. The SVM algorithm shows completely different results from others. It had higher data variation between actual and predicted WQIs.

Based on the model errors, this study found three based (DT and ExT) and ensemble tree based (XGB and RF) are more robust and reliable than other algorithms. These algorithms perform better when predicting WQIs for the coastal water quality. The finding are in line with those observed in earlier studies (Bui et al., 2020; Ghorbani et al., 2018; Khosravi et al., 2018; Kouadri et al., 2021). As seen from Fig. 13 below, DT and ExT models were predicted WQIs correctly at each monitoring sites whereas small variation was found between predicted and actual WQIs for the XGB and RF models. The results of the predictive WQIs indicate that DT and ExT models had overfitting problems because tree-based models are developed based on a single tree without any control (Biebler et al., 2009; Ying, 2019).

4.5. Justification of prediction errors of various model

To justify the model bias in predicting WQIs, the Tylor diagram analysis was utilized in this research. Recently, this approach is widely used to compare various methods/datasets/models in terms of data variances (Kärnä and Baptista, 2016; Xu et al., 2016). Fig. 13 provides an insight into how the model performed in terms of three statistical measures, where, statistics were obtained from various ML techniques by using actual and predicted WQI values in Cork Harbour respectively.

As can be seen from Fig. 14, a significant statistical difference was found among ML models with $p < 0.05$. As shown in Fig. 10, it can be clearly defined that the SVM model significantly differed from the other ML algorithms. Comparatively, the lowest RMSD, SD, and higher correlation were found for the XGB, DT and ExT algorithms respectively. In addition, relatively the RF algorithms shows well performance compare than remaining algorithms.

However, the present study compared various ML algorithms using the cross-validation results, coefficient of determination, analysis. In

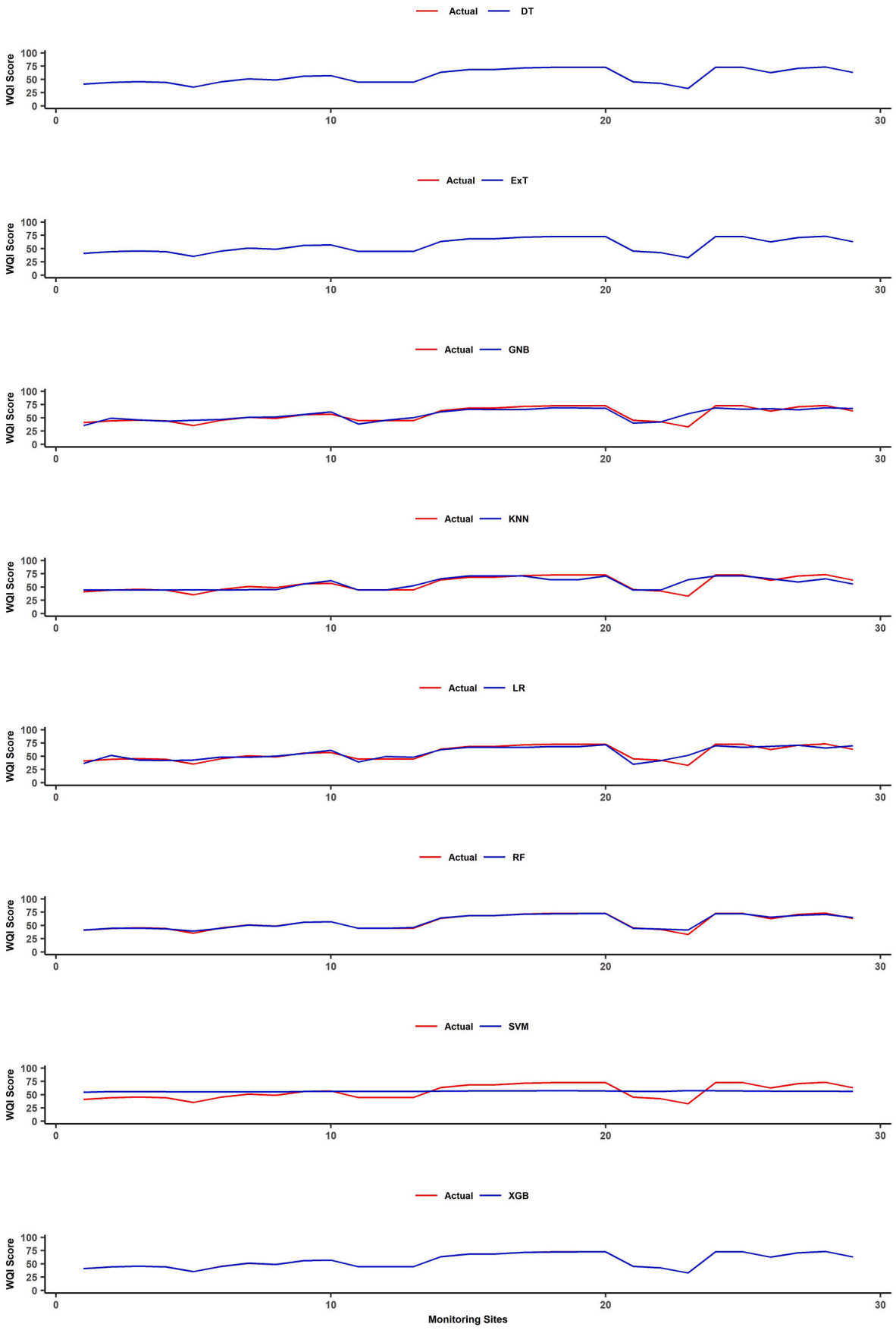

Fig. 8. Comparison of the model testing performance between predicted and actual WQI values at each monitoring sites.

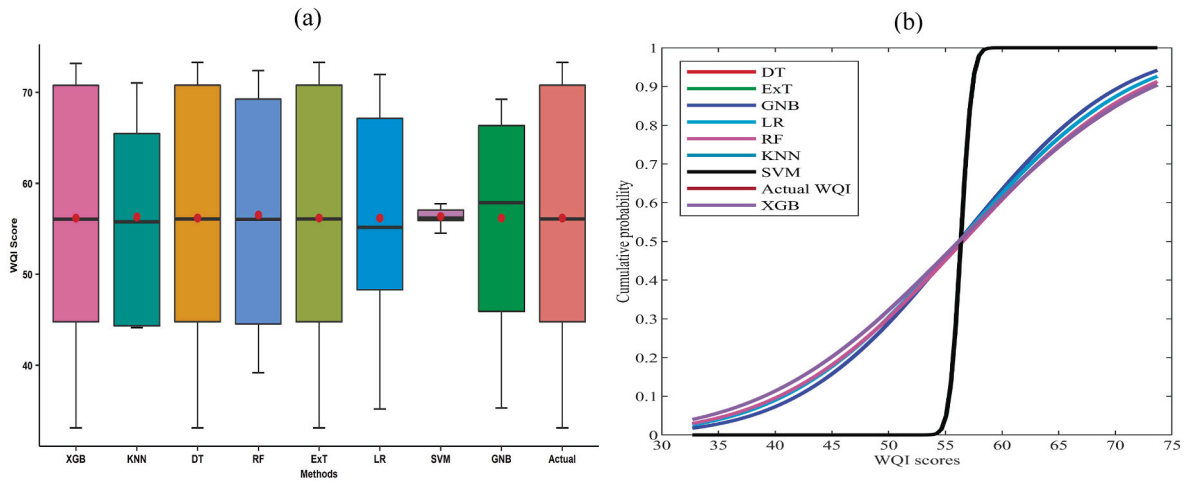

Fig. 9. Comparison of predicted WQI from various ML models: (a) Boxplots show a comparison between actual and predicted WQI scores and (b) CDF comparison of predicted WQI scores of various ML.

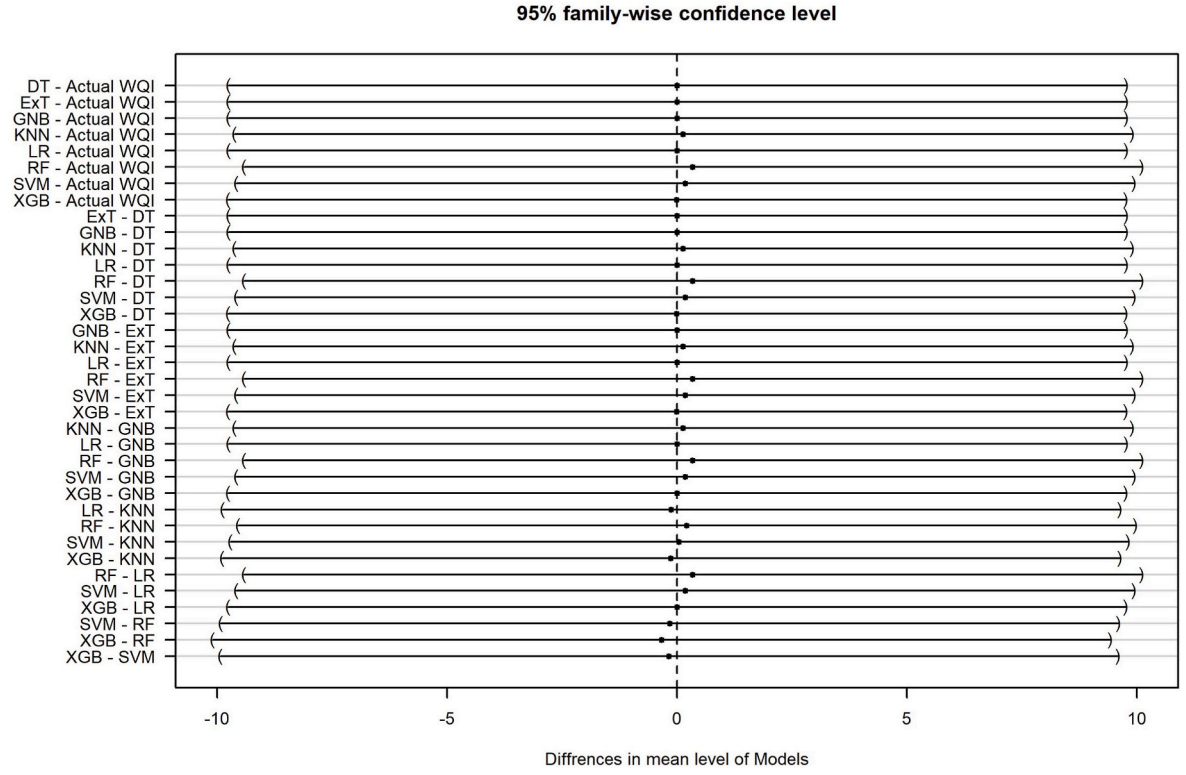

Fig. 10. Multiple comparison results of pair-wise ML models with 95% CI from Tukey's HSD, the vertical dashed line indicates the point where the difference between the means is equal to zero or similarity of model statistics, the refers to the means are equal of both models.

terms of bias between actual and predicted WQIs, the results of Tukey's HSD test, and Tylor diagram reveal that there were no statistically significant differences between predicted and actual WQI values for all algorithms except SVM. In order to evaluate the predictive performance, the CV results indicate that the XGB, DT, and ExT algorithms had the lowest prediction errors compared to other models.

Therefore, the results of this research reveal that the DT, ExT, XGB, and RF models might be effective and robust for predicting coastal WQIs in terms of reducing the WQI model uncertainty. On the other hand, it is not easy to conclude which algorithm is "better" or "worst". The present study was observed to perform well for other models except SVM.

5. Discussion

WQI model is a widely used tool to assess water quality by employing straightforward mathematical functions. Computing WQI values is relatively complex using SI and indicator weight values (Leong et al., 2019). Because, recently, several studies have revealed that these components provide a considerable uncertainty to the final assessment (Uddin et al., 2021, Uddin et al., 2022c). In this circumstance, many studies utilized the ML technique for predicting WQI values, except for SI and weight values. In this research, we used eight ML algorithms for predicting the newly developed WQM-WQI model in order to identify the most robust technique in terms of assessing coastal water quality. In

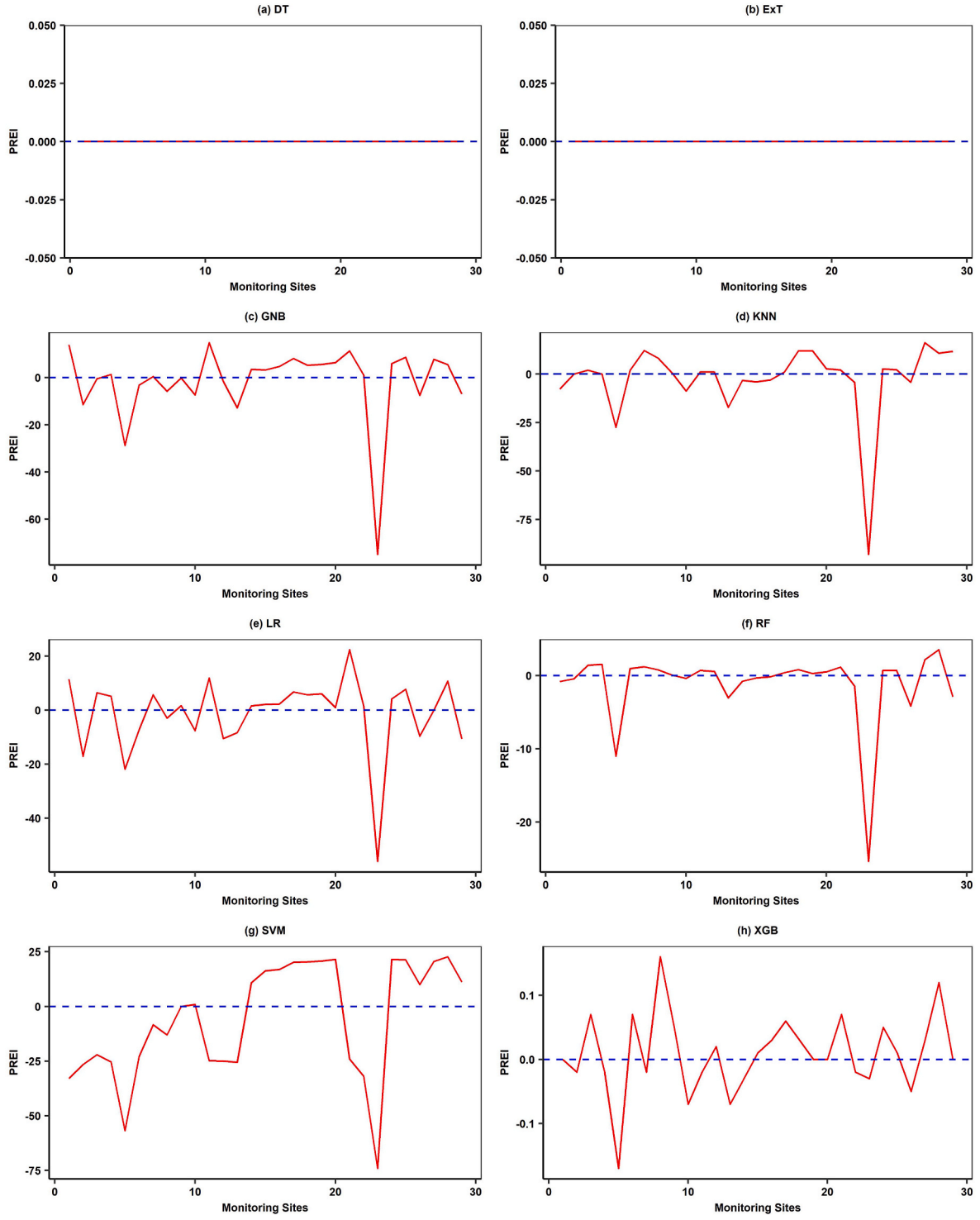

Fig. 11. WQIs predicting errors of various ML models at each monitoring sites in Cork Harbour.

this research, the WQM-WQI revealed that the water quality in Cork Harbour can be classified into two categories: "fair" and "marginal" over the study period (Table 4). Comparatively, better water quality found in the lower and the outer Harbour than in the upper part (Fig. 5). The past decade has setup an increase in the use of ETPs in this area. Recently, several annual reports of the EPA's Ireland revealed that the ETPs could contribute to raw wastewater discharges into the estuary directly without any modification of water attributes (Hartnett and Nash, 2015, Fig. 1). As a result, it is expected that the water quality in the upper part of the Harbour's associated with relatively downgraded water quality due to the extremely loaded the wastewater.

In the present study, eight widely used algorithms tested in order to identify the most robust model. Details of the prediction results are provided in Fig. 8 and Annex (e). Based on the model performance metrics, compared to the models, DT and Ext showed the outperformed capabilities to predict WQM-WQIs. However, model overfitting problems were found for these algorithms over the testing period due to the small dataset (Song and Lu, 2015; Vabalas et al., 2019). Compared with other algorithms like SVM and KNN, the DT provides better results that are effective for the prediction (Huynh-Cam et al., 2021). Unlike the

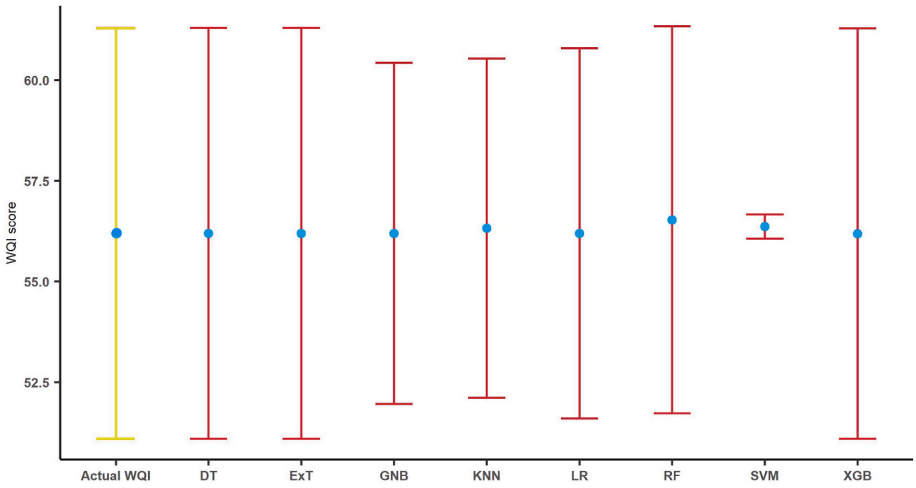

Fig. 12. Statistical significance of predicting uncertainty in WQIs for various ML models with 95% CI. Here, 95% CI where n is 29, $p < 0.0001$. Green error bar represent the measured (actual) WQIs using WQM-WQI model. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

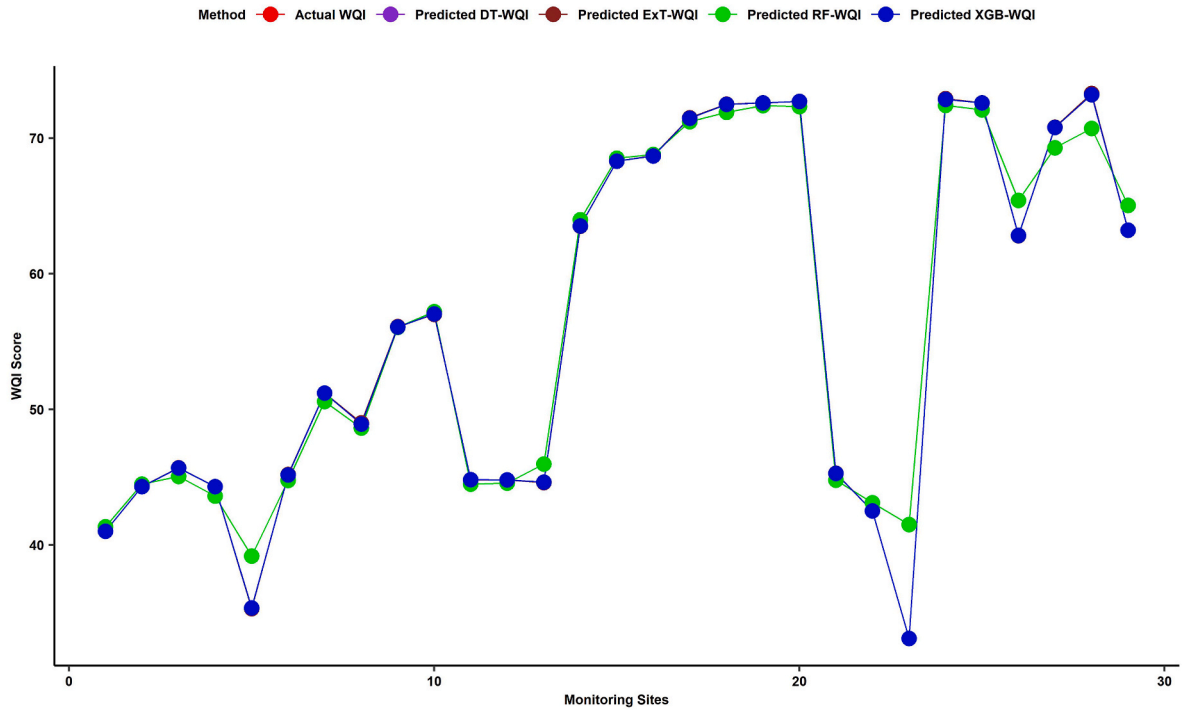

Fig. 13. Comparison among the outperformed machine learning algorithms.

decision tree model, ensemble based bagging RF and boosting XGB algorithms outperformed in predicting WQM-WQIs during both training and testing periods. Recently, several studies have also revealed that the XGB algorithm is effective for predicting WQIs (Grbčić et al., 2021; Huan et al., 2020; Islam Khan et al., 2021; Uddin et al., 2022b). Because the ensemble based algorithms combine multiple DTs and consider the average of the output of all DTs for the prediction (Malek et al., 2022). In contrast, non-parametric KNN, Gaussian based GNB, and LR algorithms showed better performance than the SVM model. In this study, the worst performance found for the SVM during both the training and testing periods (Fig. 6). Mostly, the SVM model performance is influenced by the distribution of input variables and the number of inputs (Vabalas et al., 2019). As shown in Fig. 3, model inputs DOX, DIN, TON, and CHL had a negative left skew data distribution, whereas the remaining inputs had a normal distribution. The SVM model performance was worst during the testing period due to input variation because the SVM model prediction results are highly sensitive to the significant features (Kaliappan et al., 2021; Veropoulos et al., 1999). In addition, Akbani et al. (2004) point out three causes of performance loss in the SVM prediction model: (i) positive points lying further from the ideal boundary (Wu and Chang, 2003); (ii) weakness of soft-margins (Veropoulos et al., 1999); and (iii) imbalanced support vector ratio (Wu and Chang, 2003). Details of the contorlling factors of the sensitivity of the SVM model are discussed by Akbani et al. (2004) and Veropoulos et al. (1999). In this study, the SVM model performance dropped during the testing period due to the imbalanced support vector ratio. Moreover, the lowest prediction errors (PREI) found for the DT, ExT, XGB and RF, respectively (Fig. 11). The results of this study show that the DT, EXT, XGB, and RF algorithms are efficient for predicting WQM-WQIs in terms of the lowest prediction errors when compared to other models.

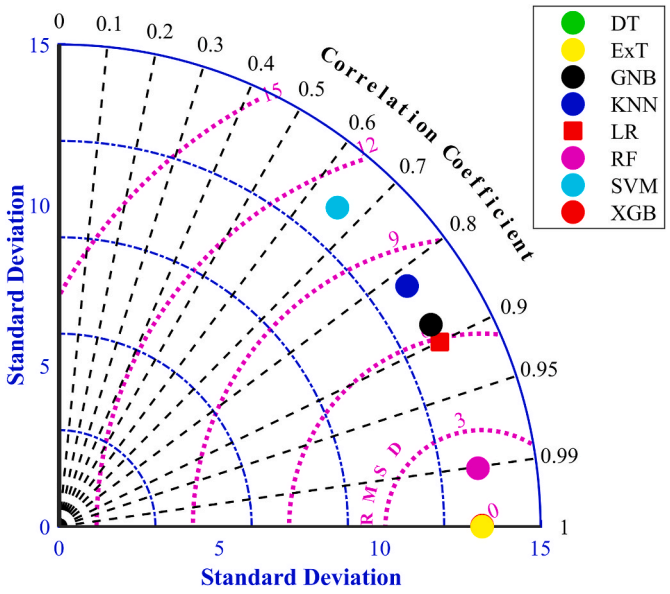

Fig. 14. Various predictive ML models comparison using Tylor diagram.

However, the findings of this research could be helpful for monitoring and assessing coastal water quality using the WQM-WQI model incorporating robust ML technique(s).

6. Conclusion

The goal of this research was to determine the robust algorithm for predicting the coastal water quality index (CWQI) accurately in terms of model uncertainty. To achieve this goal, eight ML algorithms (DT, RF, XGB, KNN, SVM, ExT, LR, and GNB) were tested and validated for predicting CWQI in Cork Harbour. Predictive models were validated using a number of validators such as RMSE, MSE, MAE, R^2 and PREI. The findings of this study can be summarized as follows:

- Compared to CV results, the XGB showed the best outperformed, whereas the lowest training (RMSE = 3.3, MSE = 10.91, MAE = 1.67, and R2 = 1.0) and testing (RMSE = 0.0, MSE = 0.0, MAE = 0.02, and R2 = 1.0) errors were found, respectively.
- The lowest prediction errors were found for the DT (PREI = 0), ExT (PREI = 0), and XGB (PERI = + 0.1 to - 0.1). They perform better in predicting WQIs at each of the monitoring sites in Cork Harbour.
- Unlike the remaining models, the RF showed better performance; its errors ranged from +1.0 to −25), whereas the remaining models had higher underestimate and overestimate problems.
- Although the Tukey's HSD family wise multi-comparison results reveal that, there were no significant difference between actual and predicted WQIs among ML models except the SVM.

Therefore, it can be concluded, based on the results of this study, that tree-based (DT and ExT) and ensemble-based (XGB and RF) algorithms could be effective and robust for predicting the CWQI. The findings of this research would also have been much more useful in predicting WQIs at each monitoring site more accurately in order to reduce the uncertainty in the WQI model. This study's inadequacy to assess the water quality in terms of temporal resolution constitutes one of its limitations. Further studies should be carried out in order to validate the other algorithms in terms of predicting WQIs using temporal variability of data attributes.

Credit author statement

Md Galal Uddin: Conceptualization, Methodology, Investigation, Formal analysis, **ML** and **AI** performed, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Stephen Nash**: Conceptualization, Methodology, Investigation, Writing – review & editing. **Mir Talas Mahammad Diganta**: Visualization, Writing – review & editing. **Azizur Rahaman**: Data curation, Methodology, statistical and **ML** analysis, Writing – review & editing. **Agnieszka I. Olbert**: Conceptualization, Methodology, Investigation, Supervision, Data curation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors gratefully acknowledge the editor's and anonymous reviewers' contributions to the improvement of this paper. This research was funded by the Hardiman Research Scholarship of the National University of Ireland Galway, which funded the first author as part of his PhD program. The authors would like to thank the Environmental Protection Agency of Ireland for providing water quality data. The authors would like to acknowledge support from MaREI, the SFI Research Centre for Energy, Climate, and Marine research. The authors also sincerely acknowledge Charles Sturt University for providing all necessary supports to this PhD project through the international co-supervision. The authors gratefully acknowledge the Eco-HydroInformatics Research Group for providing computational lab facilities at the Civil Engineering, National University of Ireland Galway.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2022.115923.

References

Abbasi, T., Abbasi, S., 2012. Water Quality Indices, Water Quality Indices. https://doi.org/10.1016/C2010-0-69472-7.
Abbasi, T., Abbasi, S.A., 2011. Water quality indices based on bioassessment: the biotic indices. J. Water Health 9 (2), 330–348. https://doi.org/10.2166/wh.2011.133.
Ahmad, Z., Rahim, N.A., Bahadori, A., Zhang, J., 2017. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. Int. J. River Basin Manag. 15, 79–87. https://doi.org/10.1080/15715124.2016.1256297.
Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. In: European Conference on Machine Learning. San Antonio, USA, pp. 39–50. https://doi.org/10.1007/978-3-540-30115-8_7.
Aldhyani, T.H.H., Al-Yaari, M., Alkahtani, H., Maashi, M., 2020. Water quality prediction using artificial intelligence algorithms. Appl. Bionics Biomechanics 2020. https://doi.org/10.1155/2020/6659314.
Aschonitis, V.G., Mastrocicco, M., Colombani, N., Salemi, E., Kazakis, N., Voudouris, K., Castaldelli, G., 2012. Assessment of the intrinsic vulnerability of agricultural land to water and nitrogen losses via deterministic approach and regression analysis. Water. Air. Soil Pollut. 223, 1605–1614. https://doi.org/10.1007/s11270-011-0968-5.
Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., Aziz, F., 2021. Machine learning algorithms for efficient water quality prediction. Model. Earth Syst. Environ. https://doi.org/10.1007/s40808-021-01266-6.
Babbar, R., Babbar, S., 2017. Predicting river water quality index using data mining techniques. Environ. Earth Sci. 76, 1–15. https://doi.org/10.1007/s12665-017-6845-9.
Biebler, K.E.E., Jäger, B.P., Wodney, M., 2009. Basic Principles of Data Mining, Social Implications of Data Mining and Information Privacy: Interdisciplinary Frameworks and Solutions. https://doi.org/10.4018/978-1-60566-196-4.ch015.
Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N., 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Sci. Total Environ. 721, 137612 https://doi.org/10.1016/j.scitotenv.2020.137612.
Calim, M.C., Nobre, P., Oke, P., Schiller, A., Siqueira, L.S.P., Castelão, G.P., 2018. A new tool for model assessment in the frequency domain – spectral Taylor Diagram : application to a global ocean general circulation model with tides. Geosci. Model Dev. (GMD). https://doi.org/10.5194/gmd-2018-5.

Chang, N., Luo, L., Wang, X.C., Song, J., Han, J., Ao, D., 2020. A novel index for assessing the water quality of urban landscape lakes based on water transparency. Sci. Total Environ. 735, 139351 https://doi.org/10.1016/j.scitotenv.2020.139351.
Chiang, L.C., Wang, Y.C., Chen, Y.K., Liao, C.J., 2021. Quantification of land use/land cover impacts on stream water quality across Taiwan. J. Clean. Prod. 318, 128443 https://doi.org/10.1016/j.jclepro.2021.128443.
Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623. https://doi.org/10.7717/peerj-cs.623.
Cumming, G., Fidler, F., Vaux, D.L., 2007. Error bars in experimental biology. J. Cell Biol. 177, 7–11. https://doi.org/10.1083/jcb.200611141.
Deng, T., Duan, H.-F., Keramat, A., 2022. Spatiotemporal characterization and forecasting of coastal water quality in the semi-enclosed Tolo Harbour based on machine learning and EKC analysis. Eng. Appl. Comput. Fluid Mech. 16, 694–712. https://doi.org/10.1080/19942060.2022.2035257.
Elgeldawi, E., Sayed, A., Galal, A.R., Zaki, A.M., 2021. Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. Informatics 8, 1–21. https://doi.org/10.3390/informatics8040079.
EPA, 2017. Urban Waste Water Treatment. Official Journal of the European Union.
Fogarty, P., 2017. Protecting Our Ocean's Wealth-A Proposal for Legal Protection of Threatened Marine Species (Dublin).
Gao, Y., Qian, H., Ren, W., Wang, H., Liu, F., Yang, F., 2020. Hydrogeochemical characterization and quality assessment of groundwater based on integrated-weight water quality index in a concentrated urban area. J. Clean. Prod. 260, 121006 https://doi.org/10.1016/j.jclepro.2020.121006.
Ghorbani, M.A., Deo, R.C., Karimi, V., Yaseen, Z.M., Terzi, O., 2018. Implementation of a hybrid MLP-FFA model for water level prediction of Lake Egirdir, Turkey. Stoch. Environ. Res. Risk Assess. 32, 1683–1697. https://doi.org/10.1007/s00477-017-1474-0.
Gikas, G.D., Sylaios, G.K., Tsihrintzis, V.A., Konstantinou, I.K., Albanis, T., Boskidis, I., 2020. Comparative evaluation of river chemical status based on WFD methodology and CCME water quality index. Sci. Total Environ. 745, 140849 https://doi.org/10.1016/j.scitotenv.2020.140849.
Grbčić, L., Družeta, S., Mauša, G., Lipić, T., Lušić, D.V., Alvir, M., Lučin, I., Sikirica, A., Davidović, D., Travaš, V., Kalafatović, D., Pikelj, K., Fajković, H., Holjević, T., Kranjčević, L., 2021. Coastal Water Quality Prediction Based on Machine Learning with Feature Interpretation and Spatio-Temporal Analysis.
Gupta, S., Gupta, S.K., 2021. A critical review on water quality index tool: Genesis, evolution and future directions. Ecol. Inform. 63, 101299 https://doi.org/10.1016/j.ecoinf.2021.101299.
Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A., 2018a. Water quality prediction using machine learning methods. Water Qual. Res. J. Can. 53, 3–13. https://doi.org/10.2166/wqrj.2018.025.
Haghiabi, A.H., Nasrolahi, A.H., Parsaie, A., 2018b. Water quality prediction using machine learning methods. Water Qual. Res. J. 53, 3–13. https://doi.org/10.2166/wqrj.2018.025.
Hartnett, M., Nash, S., 2015. An integrated measurement and modeling methodology for estuarine water quality management. Water Sci. Eng. 8, 9–19. https://doi.org/10.1016/j.wse.2014.10.001.
Hassan, Md, Mehedi, Hassan, Mahedi, Md, Akter, L., Rahman, M.M., Zaman, S., Hasib, K. M., Jahan, N., Smrity, R.N., Farhana, J., Raihan, M., Mollick, S., 2021. Efficient prediction of water quality index (WQI) using machine learning algorithms. Human-Centric Intell. Syst. 1, 86. https://doi.org/10.2991/hcis.k.211203.001.
He, X., Gou, W., Liu, Y., Gao, Z., 2015. A practical method of nonprobabilistic reliability and parameter sensitivity analysis based on space-filling design. Math. Probl Eng. 1–12. https://doi.org/10.1155/2015/561202, 2015.
Huan, J., Li, H., Li, M., Chen, B., 2020. Prediction of dissolved oxygen in aquaculture based on gradient boosting decision tree and long short-term memory network: a study of Chang Zhou fishery demonstration base, China. Comput. Electron. Agric. 175, 105530 https://doi.org/10.1016/j.compag.2020.105530.
Huynh-Cam, T.T., Chen, L.S., Le, H., 2021. Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. Algorithms 14. https://doi.org/10.3390/a14110318.
Ireland, 2001. Parameters of water quality. Environmental Protection Agency, Ireland. https://doi.org/10.1017/CBO9781107415324.004.
Islam Khan, M.S., Islam, N., Uddin, J., Islam, S., Nasir, M.K., 2021. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. J. King Saud Univ. - Comput. Inf. Sci. 34 (8), 4773–4781. https://doi.org/10.1016/j.jksuci.2021.06.003.
Juwana, I., Muttil, N., Perera, B.J.C., 2016. Uncertainty and sensitivity analysis of west java water sustainability index - a case study on citarum catchment in Indonesia. Ecol. Indicat. 61, 170–178. https://doi.org/10.1016/j.ecolind.2015.08.034.
Kadam, A.K., Wagh, V.M., Muley, A.A., Umrikar, B.N., Sankhua, R.N., 2019. Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. Model. Earth Syst. Environ. 5, 951–962. https://doi.org/10.1007/s40808-019-00581-3.
Kaliappan, J., Srinivasan, K., Mian Qaisar, S., Sundararajan, K., Chang, C.Y., Suganthan, C., 2021. Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. Front. Public Health 9, 1–12. https://doi.org/10.3389/fpubh.2021.729795.
Kärnä, T., Baptista, A.M., 2016. Evaluation of a long-term hindcast simulation for the Columbia River estuary. Ocean Model. 99, 1–14. https://doi.org/10.1016/j.ocemod.2015.12.007.
Khan, Y., See, C.S., 2016. Predicting and analyzing water quality using Machine Learning: a comprehensive model. In: 2016 IEEE Long Isl. Syst. Appl. Technol. Conf. LISAT 2016. https://doi.org/10.1109/LISAT.2016.7494106.
Khosravi, K., Mao, L., Kisi, O., Yaseen, Z.M., Shahid, S., 2018. Quantifying hourly suspended sediment load using data mining models: case study of a glacierized Andean catchment in Chile. J. Hydrol. 567, 165–179. https://doi.org/10.1016/j.jhydrol.2018.10.015.
Khullar, S., Singh, N., 2021. Machine learning techniques in river water quality modelling: a research travelogue. Water Sci. Technol. Water Supply 21. https://doi.org/10.2166/ws.2020.277.
Kouadri, S., Elbeltagi, A., Islam, A.R.M.T., Kateb, S., 2021. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl. Water Sci. 11, 1–20. https://doi.org/10.1007/s13201-021-01528-9.
Leong, W.C., Bahadori, A., Zhang, J., Ahmad, Z., 2019. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). Int. J. River Basin Manag. 1–8. https://doi.org/10.1080/15715124.2019.1628030, 0.
Malek, N.H.A., Wan Yaacob, W.F., Md Nasir, S.A., Shaadan, N., 2022. Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water. https://doi.org/10.3390/w14071067.
Mohammed, H., Longva, A., Seidu, R., 2018. Predictive analysis of microbial water quality using machine-learning algorithms. Environ. Res. Eng. Manag. 74, 7–20. https://doi.org/10.5755/j01.erem.74.1.20083.
Nash, S., Hartnett, M., Dabrowski, T., 2011. Modelling phytoplankton dynamics in a complex estuarine system. Proc. Inst. Civ. Eng. - Water Manag. 164, 35–54. https://doi.org/10.1680/wama.800087.
Othman, F., Alaaeldin, M.E., Seyam, M., Ahmed, A.N., Teo, F.Y., Ming Fai, C., Afan, H.A., Sherif, M., Sefelnasr, A., El-Shafie, A., 2020. Efficient river water quality index prediction considering minimal number of inputs variables. Eng. Appl. Comput. Fluid Mech. 14, 751–763. https://doi.org/10.1080/19942060.2020.1760942.
Pham, B.T., Prakash, I., Khosravi, K., Chapi, K., Trinh, P.T., Ngo, T.Q., Hosseini, S.V., Bui, D.T., 2019. A comparison of Support Vector Machines and Bayesian algorithms for landslide susceptibility modelling. Geocarto Int. 34, 1385–1407. https://doi.org/10.1080/10106049.2018.1489422.
Prasad, D.V.V., Venkataramana, L.Y., Kumar, P.S., Prasannamedha, G., Harshana, S., Srividya, S.J., Harrinei, K., Indraganti, S., 2022. Analysis and prediction of water quality using deep learning and auto deep learning techniques. Sci. Total Environ. 821, 153311 https://doi.org/10.1016/j.scitotenv.2022.153311.
Prakash, R., Tharun, V.P., Renuga Devi, S., 2018. A Comparative Study of Various Classification Techniques to Determine Water Quality. Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 1501–1506. https://doi.org/10.1109/ICICCT.2018.8473168.
Rahman, A., 2020. Statistics for Data Science and Policy Analysis, Statistics for Data Science and Policy Analysis. Springer Singapore. https://doi.org/10.1007/978-981-15-1735-8.
Rahman, A., 2019. Statistics-based data preprocessing methods and machine learning algorithms for big data analysis | rahman | international journal of artificial Intelligence. Int. J. Artif. Intell. 17, 44–65.
Rahman, A., Harding, A., 2016. Small area estimation and microsimulation modeling. In: Small Area Estimation and Microsimulation Modeling, first ed. CRC Press https://doi.org/10.1201/9781315372143/SMALL-AREA-ESTIMATION-MICROSIMULATION-MODELING-AZIZUR-RAHMAN-ANN-HARDING.
Rezaie-Balf, M., Attar, N.F., Mohammadzadeh, A., Murti, M.A., Ahmed, A.N., Fai, C.M., Nabipour, N., Alaghmand, S., El-Shafie, A., 2020. Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: comparative assessment of a noise suppression hybridization approach. J. Clean. Prod. 271 https://doi.org/10.1016/j.jclepro.2020.122576.
Seifi, A., Dehghani, M., Singh, V.P., 2020. Uncertainty analysis of water quality index (WQI) for groundwater quality evaluation: application of Monte-Carlo method for weight allocation. Ecol. Indicat. 117, 106653 https://doi.org/10.1016/j.ecolind.2020.106653.
Sharif, O., Hasan, M.Z., Rahman, A., 2022. Determining an effective short term COVID-19 prediction model in ASEAN countries. Sci. Rep. 12, 1–11. https://doi.org/10.1038/s41598-022-08486-5.
Smith, D.G., 1990. A better water quality indexing system for rivers and streams. Water Res. 24, 1237–1244. https://doi.org/10.1016/0043-1354(90)90047-A.
Solanki, A., Agrawal, H., Khare, K., 2015. Predictive analysis of water quality parameters using deep learning. Int. J. Comput. Appl. 125, 29–34. https://doi.org/10.5120/ijca2015905874.
Song, Y.Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27, 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.
Stoner, J.D., 1978. Water-quality indices for specific water uses. Geol. Surv. Circular. https://doi.org/10.3133/cir770.
Shekar, B.H., Dagnew, G., 2019. Grid search-based hyperparameter tuning and classification of microarray cancer data. 2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP. https://doi.org/10.1109/ICACCP.2019.8882943.
Sutadian, A.D. hany, Muttil, N., Yilmaz, A.G. okhan, Perera, B.J.C., 2016. Development of river water quality indices-a review. Environ. Monit. Assess. 188, 58. https://doi.org/10.1007/s10661-015-5050-0.
Uddin, G., Nash, S., Olbert, A.I., 2021. A review of water quality index models and their use for assessing surface water quality. Ecol. Indicat. 122, 107218 https://doi.org/10.1016/j.ecolind.2020.107218.
Uddin, M.G., Moniruzzaman, M., Quader, M.A., Hasan, M.A., 2018. Spatial variability in the distribution of trace metals in groundwater around the Rooppur nuclear power

plant in Ishwardi, Bangladesh. Groundw. Sustain. Dev. 7 https://doi.org/10.1016/j.gsd.2018.06.002.

Uddin, M.G., Nash, Stephen, Olbert, A.I., 2020. Application of water quality index models to an Irish estuary. In: Kieran Runae, V.J. (Ed.), Civil and Environmental Research. Civil and Environmental Research, Cork, Ireland, pp. 576–581.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022. A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment. Water Res. 219, 118532 https://doi.org/10.1016/J.WATRES.2022.118532.

Uddin, G., Nash, S., Olbert, A.I., 2022a. Optimization of Parameters in a Water Quality Index Model Using Principal Component Analysis.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2022b. Development of a water quality index model - a comparative analysis of various weighting methods. In: Çiner, P.D.A. (Ed.), Mediterranean Geosciences Union Annual Meeting (MedGU-21). Istanbul, pp. 1–6.

Uddin, M.G., Nash, S., Diganta, M.T.M., Rahman, A., Olbert, A.I., 2022c. A comparison of geocomputational models for validating geospatial distribution of water quality index. In: Priyanka, H., Rahman, A., Basant agarwal, Binita Tiwari (Eds.), Computational Statistical Methodologies and Modeling for Artificial Intelligence. CRC Press, Taylor & Francis Publisher, USA.

Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. PLoS One 14, e0224365. https://doi.org/10.1371/journal.pone.0224365.

Veropoulos, K., Campbell, C., Cristianini, N., Others, 1999. Controlling the sensitivity of support vector machines. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 55–60.

Venkata Vara Prasad, V.V.P., Y Venkataramana, L., Kumar, P.S., G, P., K, S., A.J, P, 2020. Water quality analysis in a lake using deep learning methodology: prediction and validation. Int. J. Environ. Anal. Chem. https://doi.org/10.1080/03067319.2020.1801665.

Villalobos-Arias, L., Quesada-López, C., Guevara-Coto, J., Martínez, A., Jenkins, M., 2020. Evaluating hyper-parameter tuning using random search in support vector machines for software effort estimation. In: PROMISE 2020 - Proc. 16th ACM Int. Conf. Predict. Model. Data Anal. Softw. Eng. Co-located with ESEC/FSE 2020, pp. 31–40. https://doi.org/10.1145/3416508.3417121.

Wang, X., Zhang, F., Ding, J., 2017. Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. Sci. Rep. 7, 1–18. https://doi.org/10.1038/s41598-017-12853-y.

Wu, G., Chang, E.Y., 2003. Class-boundary alignment for imbalanced dataset learning. ICML Work. Learn. from Imbalanced Data Sets II 49–56.

Wu, J., Wang, Z., 2022. A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory, vol. 14. https://doi.org/10.3390/w14040610. Water (Switzerland).

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., Hu, J., 2020. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput. Mater. Sci. 171, 109203 https://doi.org/10.1016/j.commatsci.2019.109203.

Xu, Z., Hou, Z., Han, Y., Guo, W., 2016. A diagram for evaluating multiple aspects of model performance in simulating vector fields. Geosci. Model Dev. (GMD) 9, 4365–4380. https://doi.org/10.5194/gmd-9-4365-2016.

Yan, F., Qiao, D., Qian, B., Ma, L., Xing, X., Zhang, Y., Wang, X., 2016. Improvement of CCME WQI using grey relational method. J. Hydrol. 543, 316–323. https://doi.org/10.1016/j.jhydrol.2016.10.007.

Ying, X., 2019. An overview of overfitting and its solutions. J. Phys. Conf. Ser. 1168 https://doi.org/10.1088/1742-6596/1168/2/022022.

Zotou, I., Tsihrintzis, V.A., Gikas, G.D., 2019. Performance of Seven Water Quality Indices (WQIs) in a Mediterranean River. Environ. Monit. Assess. 191 https://doi.org/10.1007/s10661-019-7652-4.