Journal of Cleaner Production 450 (2024) 141877

Contents lists available at ScienceDirect

Journal of Cleaner Production

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

Monitoring the Industrial waste polluted stream - Integrated analytics and machine learning for water quality index assessment

Ujala Ejaz [a,b], Shujaul Mulk Khan [a,*], Sadia Jehangir [a,b], Zeeshan Ahmad [c,***], Abdullah Abdullah [a], Majid Iqbal [a,d], Noreen Khalid [e], Aisha Nazir [f], Jens-Christian Svenning [b,**]

[a] Department of Plant Sciences, Quaid-i-Azam University, Islamabad, 45320, Pakistan
[b] Center for Ecological Dynamics in a Novel Biosphere (ECONOVO), Department of Biology, Aarhus University, Ny Munkegade 114, DK-8000, Aarhus C, Denmark
[c] CAS Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengl, 666303, China
[d] Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, China
[e] Department of Botany, Government College Women University, Sialkot, Pakistan
[f] Environmental Biotechnology Research Laboratory, Institute of Botany, University of the Punjab, Lahore, Pakistan

ARTICLE INFO

Handling editor: Jin-Kuk Kim

Keywords:
Industrial wastewater
Machine learning
Assessment and monitoring
Water Quality Index
Artificial Intelligence

ABSTRACT

The Water Quality Index (WQI) is a primary metric used to evaluate and categorize surface water quality which plays a crucial role in the management of fresh water resources. Machine Learning (ML) modeling offers potential insights into water quality index prediction. This study employed advanced ML models to get potential insights into the prediction of water quality index for the Aik-Stream, an industrially polluted natural water resource in Pakistan with 19 input water quality variables aligning them with surrounding land use and anthropogenic activities. Six machine learning algorithms, i.e. Adaptive Boosting (AdaBoost), K-Nearest Neighbors (K-NN), Gradient Boosting (GB), Random Forests (RF), Support Vector Regression (SVR), and Bayesian Regression (BR) were employed as benchmark models to predict the Water Quality Index (WQI) values of the polluted stream to achieve our objectives. For model calibration, 80% of the dataset was reserved for training, while 20% was set aside for testing. In our comparative analyses of predictive models for water quality index, the Gradient Boost (GB) model stood out the fittest for its precision, utilizing a combination of just seven parameters (chemical oxygen demand, total organic carbon, oil & grease, Ammonia- nitrogen, arsenic, nickel and zinc), surpassing other models by achieving better results in both training ($R^2 = 0.88$, RMSE = 7.24) and testing ($R^2 = 0.85$, RMSE = 8.67). Analyzing feature importance showed that all the selected variables, except for NO_3 N, TDS and temperature had an impact on the accuracy of the models predictions. It is concluded that the application of machine learning to assess water quality in polluted environments enhances accuracy and facilitates real-time tracking, enabling proactive risk mitigations.

1. Introduction

Water pollution from industries is one of the major global problems, especially in rapidly developing countries, where numerous factories often release their effluents directly into the nearby water tributaries (Zhang et al., 2021). Industrial waste products contaminate water with highly toxic metals and organic pollutants (Teo et al., 2022; Whitehead et al., 2018). The rate of global industrial wastewater release is projected to be doubled by 2025 (Hutton and Shafahi, 2019; Water, 2017), an escalation which could pose a major threat to the freshwater resources. The Lancet Commission on Pollution and Health identified water pollution as the leading cause of premature deaths worldwide (Landrigan et al., 2017). The UN Sustainable Development Goals (SDGs) also aim to improve water quality by minimizing the emission of hazardous chemicals and enhancing recycling and safe reuse by the end of 2030 (Weiland et al., 2021). Pakistan is ranked 17th in the world in terms of acute water scarcity, with 79% of its water unsafe for drinking (Jabeen et al., 2015). Only 1% of industrial wastewater is treated before being

* Corresponding author. Department of Plant Sciences, Quaid-i-Azam University, Islamabad, 45320, Pakistan.
** Corresponding author.
*** Corresponding author
E-mail addresses: smkhan@qau.edu.pk (S.M. Khan), zeeshanahmad@xtbg.ac.cn (Z. Ahmad), svenning@bio.au.dk (J.-C. Svenning).

https://doi.org/10.1016/j.jclepro.2024.141877
Received 17 November 2023; Received in revised form 19 March 2024; Accepted 22 March 2024
Available online 28 March 2024

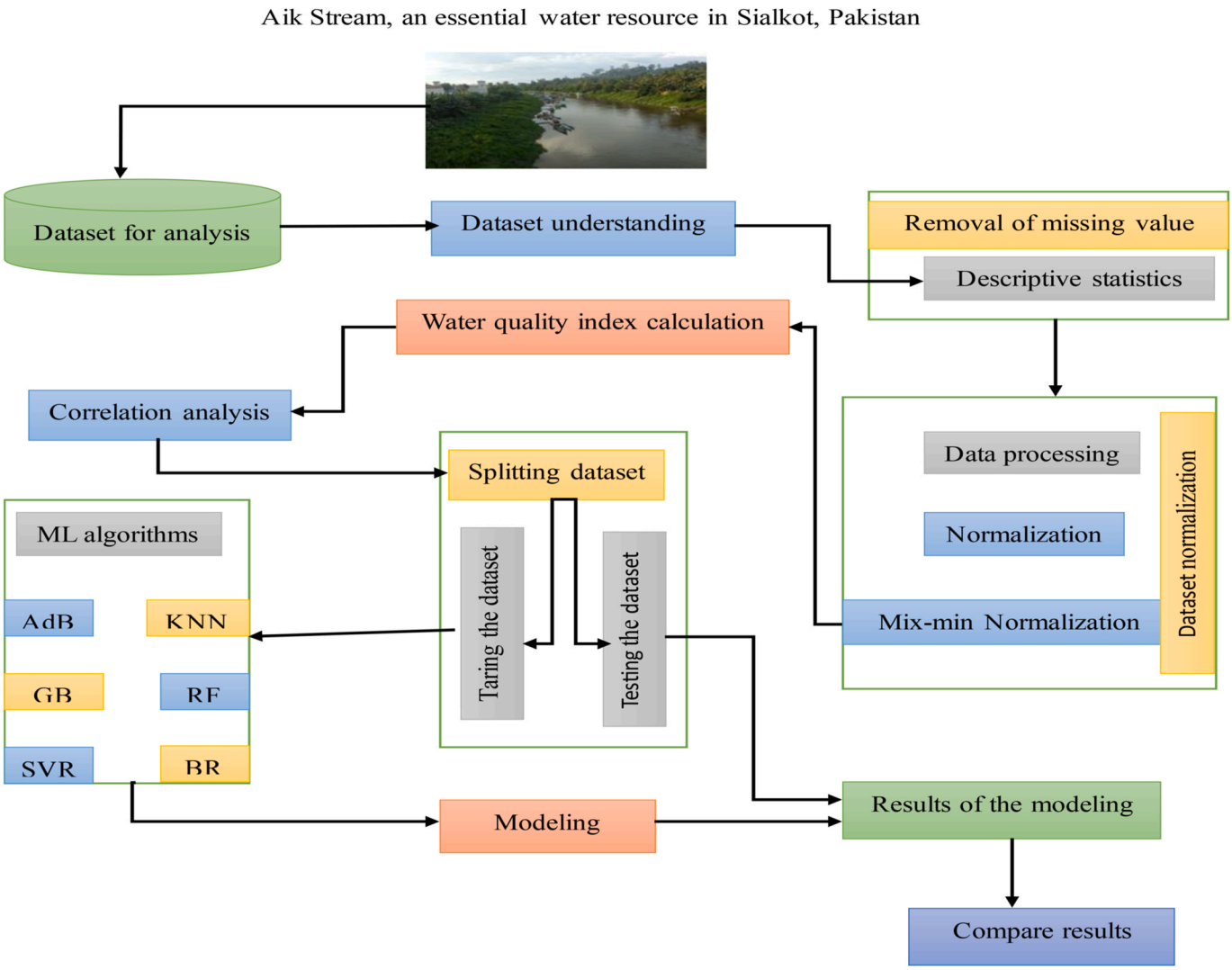

Fig. 1. Proposed working model diagram.

discharged directly into rivers and drains (Fida et al., 2022). Wastewater produced by industries here is often discharged into nearby water bodies without treatment (Bhagat et al., 2023). Streams and rivers have been turned into open sewers due to the discharge from heavily industrialized, densely populated areas (Issakhov et al., 2021; Wang et al., 2022). The situation of the Aik-Stream, an important India-Pakistan Transboundary tributary of the Chenab River, crossing through the industrial area of Sialkot, Pakistan is the focus in this study. Over the past decade, Sialkot city has undergone rapid industrialization and urbanization making it highly susceptible to environmental pollution (Khalid et al., 2021). The city is renowned globally for its production of leather goods, sports equipments, especially quality foot balls, processed food items, ceramics and surgical instruments (Qadir et al., 2008). It houses approximately 92 tanneries, 244 units for leather garment and product manufacturing, more than 900 factories that manufacture sports goods along, with 57 units dedicated to husking rice and 14 mills that produce flour (Qadir and Malik, 2009). Unfortunately, there is inadequate disposal of municipal and industrial waste in the region, leading to the unregulated discharge of solid waste and effluents from industries directly into the stream. Over time, these pollutants haave increasingly damaged the ecological integrity of the stream, primarily due to untreated waste from leather industries (Naeem et al., 2021). Waste discharge from the leather industries includes organic and inorganic substances, toxic materials, i.e., heavy metals, chemically synthesized tannins, oils, resins, bio-toxins, and disinfectants (Garai, 2014; Maqbool et al., 2018; Rabelo et al., 2018; Tariq et al., 2010).

Continuous monitoring is essential for regulating surface water quality, effectively and it is crucial for ecosystem protection, human health, sustainable water resource management, pollution control, and policy development (Brack et al., 2017; Geissen et al., 2015). Monitoring and assessing water quality can enable the identification of potential risks, the mitigation of health risks, the preservation of natural environments, and the assurance of the availability of clean water for present and future generations (Mokarram et al., 2020). However, relying solely on conventional monitoring methods is inadequate. Instead, it is often essential to employ mechanisms or models to predict potential risks in less time to prevent negative impacts on water quality (Islam et al., 2021). Based on these insights, one can take several measures including initiating treatments to counteract contaminants, diversifying or switching water sources, issuing timely public advisories for community safety, adjusting and enhancing water treatment infrastructure, and implementing emergency response protocols tailored to specific threats (Berglund et al., 2020; Sun and Scanlon, 2019).

The prediction of water quality is of paramount importance for various socio-economic sectors that heavily depend on access to clean and safe water resources. More recently, artificial intelligence (AI) has become a viable field to work in when it comes to creating advanced algorithms and prediction methods that can be used to estimate the state

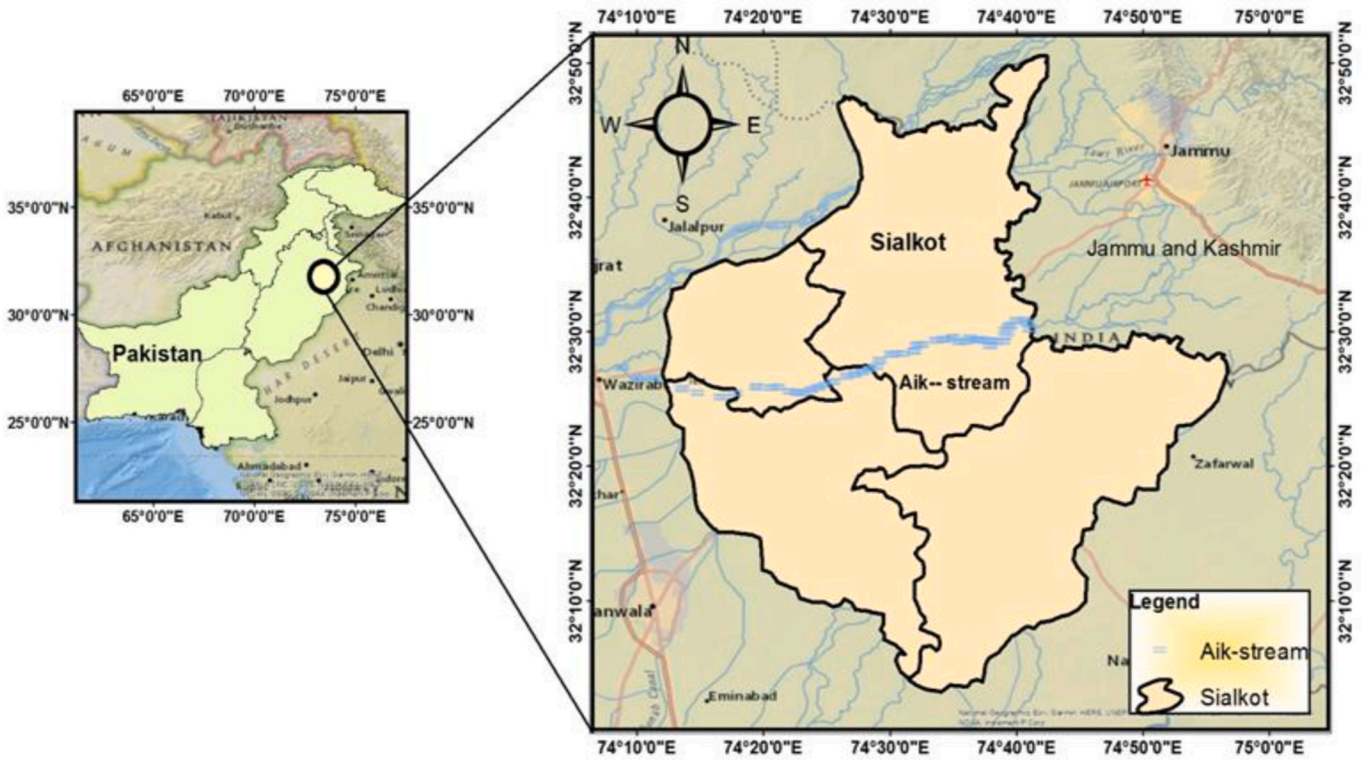

Fig. 2. The study area including the Aik stream adjacent to the industrialized city of Sialkot in Pakistan (Source: Arc map 10.5).

of water quality by analyzing complex data (Ahmed et al., 2024). This work intends to give a general overview of the major approaches used in AI-based water quality prediction, emphasizing six machine learning algorithms in particular: Adaptive Boosting (AdaBoost), Random Forests (RF), Gradient Boosting (GB), Support Vector Regression (SVR), and Bayesian Regression (BR). The aim is to improve the accuracy of water quality predictions while accounting for the complexity of the polluted stream dataset (Bhagat et al., 2023). These models may effectively predict water quality by using various monitoring techniques, managing complex relationships, supporting various types of data, and quantifying imprecision in forecasts associated with water resource management (Xu et al., 2022). Ensemble techniques such as Random Forest, Gradient Boosting, and Adaptive Boosting are widely recognized for their ability to understand complex data correlations by combining predictions from several models (Huan and Liu, 2024). This enhances performance, especially in scenarios involving intricate, non-linear patterns. (Mienye and Sun, 2022). In contrast, K-Nearest Neighbors (K-NN) stands out as a versatile algorithm capable of adapting to various data types, making it an adaptable choice for diverse datasets (Modaresi and Araghinejad, 2014). SVR recognized by constructing hyperplanes in high-dimensional spaces, enabling precise representation of intricate patterns (Liu et al., 2013). Simultaneously, Bayesian Regression offers a unique approach by incorporating prior knowledge and leveraging probabilistic modeling techniques, not only providing insights into the data but also quantifying prediction uncertainty, which proves valuable for nuanced decision-making (Sharma and Goyal, 2017). Considering recent research findings, it has been observed that the Support Vector Machine (SVM), random forest, KNN, and XGBoost algorithms have proven to be efficient in predicting water quality (Danades et al., 2016; Khan et al., 2022). Similarly, Bayesian Regression (BR) has also demonstrated effectiveness in water quality prediction, as indicated by previous studies (Li et al., 2021). The best suited model can be assessed or identified for predicting water quality.

This study compared six different machine learning methods, including AdaBoost, K-Nearest Neighbors, Gradient Boosting, Random Forests, Support Vector Regression, and Bayesian Regression, to predict the water quality index (WQI) based on the water quality parameters of Aik-Stream, Sialkot, Pakistan. Our research objectives include: (1) Assessing the surface Water Quality Index (WQI) and its spatial variation along the Aik-Stream, (2) Constructing predictive AI models and conducting comparative analyses among them, and (3) Identifying the most effective models for predicting surface WQI.

The selection of suitable model inputs was based on criteria that prioritized the lowest root mean square error (RMSE) and the highest coefficient of determination (R^2). This research methodology shares similar objectives with the studies conducted by (Gazzaz et al., 2012; Hameed et al., 2017). However, it introduces a novel approach involving the use of standardized regression coefficients to evaluate the most influential independent parameters across all six predictive models. This approach enhances our understanding of the key variables that significantly impact the Water Quality Index (WQI) in the industrially disturbed Aik-Stream of Pakistan. This innovative method not only helps fill gaps in existing research but also uses machine learning models to study water quality in polluted areas, thereby presenting novel insights pertinent to the domains of pollution management and sustainability.

2. Materials & methods

The overall approach used in the reported study included data collection, data interpretation and application of various ML models, integrated in a coordinated workflow (Fig. 1).

2.1. Study area

The Aik-Stream, a significant tributary of the Chenab River flowing through the city of Sialkot, Pakistan. Sialkot city is situated within a humid subtropical climatic zone with an average annual precipitation rate of 957.9 mm or 37.7 inches (Ali et al., 2020). The stream starts from the Pir Punjal Range in the Lesser Himalayan region of the Jammu and Kashmir, at an elevation of 530 m at sea level (Malik et al., 2010)

Table 1
Classification of CCME-WQI and corresponding water status.

Water Quality Class	Index Range	Water Status
Excellent	95–100	Very good quality water
Good	80–94	Good quality water
Fair	65–79	Acceptable quality water
Marginal	45–64	Poor quality water
Poor	0–44	Very poor-quality water

(Fig. 2). It spans a total distance of 131.6 km, with a 315 Cs annual flow rate and a catchment area covering approximately 1062 km^2. Typically, the stream receives its lowest discharge during the early summer, while the utmost occurs during the monsoon season (Mahmood et al., 2014). As the stream flows through the city, it receives a substantial influx of wastewater, including toxic chemicals and heavy metals from municipal waste and industries, which drain down into the Chenab River without treatment. This stream receives a total of 52 million liters of wastewater per day, with an additional 1.1 million units of waste from leather-producing factories (Daily, 2006; Pakistan, 2007; Qadir and Malik, 2009). The waste generated during the production of leather contains a range of pollutants, including inorganic substances. These pollutants consist of metals, created tannins, different types of oils and resins as well, as biological toxins and disinfectants (Garai, 2014; Maqbool et al., 2018; Rabelo et al., 2018; Tariq et al., 2010). Researchers have reported high levels of heavy metals such as chromium, lead, cadmium, mercury, copper, zinc and nickel from the Sialkot Industrial Zone (Jadoon and Malik, 2019; Khalid et al., 2021; Lokhande et al., 2011; Malik et al., 2010; Qadir and Malik, 2009; Qadir et al., 2008).

2.2. Dataset for analyses

The water samples were collected at 150 sampling sites in three replicates (150 × 3 samples) from the Aik-Stream at regular intervals with 1-km spacing, spanning a total of 131.6 km and ensuring comprehensive coverage of stream flow from its source to the endpoint. The samples were collected in April and at the end of September, corresponding to the hydrologically high and low flow period, respectively. Water samples were collected 30 cm below the surface within a 100-m radius of each site. Three sub-samples per site were combined into composite samples and stored in nitric acid-cleaned plastic bottles. These were sealed tightly and transported to the lab in ice boxes, following prescribed standard method (Rodier et al., 2009). One mL of high-quality nitric acid was added to each sample (to make sure its pH remained below 2) and preserve the water samples for further analysis of metals ion concentrations. The collected samples were stored at a temperature of 4 °C to prevent any degradation prior to conducting the chemical analysis, followed the guidelines set by (Association, 1926). The samples were examined to determine various important characteristics related to physicochemical and biochemical parameters. The dataset includes 19 significant explanatory variables, namely: pH, Temperature (T, ^{o}C), Electrical Conductivity (EC, μS cm^{-1}), Total Dissolved Solids (TDS, mg/L), Total Suspended Solids (TSS, mg/L), Biological Oxygen Demand (BOD, mg/L), Chemical Oxygen Demand (COD, mg/L), Ammonia Nitrogen (NH_3–N, mg/L), Total Organic Carbon (TOC, mg/L), Chloride (Cl^-, mg/L), Nitrate Nitrogen (NO_3–N, mg/L) and Oil and Grease (O&G, mg/L) zinc (Zn), nickel (Ni), copper (Cu), chromium (Cr), lead (Pb), arsenic (As), and mercury (Hg) (Table 1). Details of various analyses undertaken are mentioned in (Supplementary data, Table 1).

For the heavy metal analysis the water samples were digested on a plate using a combination of nitric and perchloric acids following the guidelines provided by (Chen and Ma, 1998). The Varian FS 240AA Fast Sequential Atomic Absorption Spectrometer were used to analyze trace metals. Metal standards were verified against Fluka's standard reference material, showing minimal deviation. For each test, the average of three replicates was recorded. The precision of these methods was demonstrated by a deviation ranging between 5% and 10%.

2.3. The Water Quality Index (WQI)

The Water Quality Index (WQI) is a key method in water quality assessment. It simplifies the assessment procedure by condensing a large amount of data on water quality into a single numerical index score (Parween et al., 2022). This study employed the WQI, which was created in 2001 by the Canadian Council of Ministers of the Environment (CCME) (CCME, 2001). It falls under the category of an open index since its computation is not predefined with respect to the variables included or standards of quality. Instead, the index requires the collection of a set of factors relevant to the water body under consideration and the evaluation's intended goals. Due to its adaptability, this index can be tailored to fit the needs of different water quality monitoring programs. However, it's crucial to remember that constant application of the same parameters and criteria is necessary for meaningful comparisons of outcomes.

Utilizing the three numerical components of scope (F1), frequency (F2), and amplitude (F3), the CCME-WQI is computed. Equation (1) represents the scope factor, from which the fraction of parameters that at least once exceed the quality norms is derived. The frequency factor (shown by equation (2)) expresses the percentage of analytical results that, when all parameters are considered, exceeds the requirements. The amplitude factor, denoted by Equation (3), calculates the discrepancy between the analytical results that fall short of the required requirements and the quality criteria. These qualities, when taken allow the WQI to offer a thorough evaluation of water quality.

$$\text{Scope}: F_1 = \left(\frac{\text{Number of failed variables}}{\text{Total number of variables}}\right) \times 100 \tag{1}$$

$$\text{Frequency}: F_2 = \left(\frac{\text{Number of failed tests}}{\text{Total number of tests}}\right) \times 100 \tag{2}$$

$$\text{Amplitude}: F3 = \frac{\text{nse}}{0.01\text{nse} + 0.01} \tag{3}$$

The calculation of Factor 3 (amplitude) involves a three-step process, which is used to determine the difference between the desired value and each concentration. These steps are represented by equations (3.1), (3.2) and (3.3) in the amplitude formulation.

$$\text{Excursion}_i = \left(\frac{\text{Failed test value}_{\ i}}{\text{Objective}_{\ j}}\right) - 1 \tag{3.1}$$

The test value should not be less than the target value in the following circumstances:

$$\text{Excursion}_i = \left(\frac{\text{Objective}_{\ j}}{\text{Failed test value}_{\ i}}\right) - 1 \tag{3.2}$$

The total deviation coefficients of all individual results, no matter whether they meet the predetermined objectives, must be added up and divided by the sum of the number of individual results to determine the overall extent of non-compliance; this value is represented by the normalized total of all the deviation coefficients (NSE), which is written as:

$$\text{NSE} = \left(\frac{\sum_{i=1}^{n} \text{Excursion}_i}{\text{Number of tests}}\right) \tag{3.3}$$

NSE from the targets is then mapped using an asymptotic function, assigning values in the range of 0–100 to calculate term F3 (Eq. (3)).

After determining these factors, the actual WQI can be computed by these three terms as vectors and incorporating them together. The sum

Table 2
Models developed from the different combinations of the water quality variables.

Models	Variables																		
1	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS	Pb	pH	T	Hg	NO_3N	Cu
2	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS	Pb	pH	T	Hg	NO_3N	
3	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS	Pb	pH	T	Hg		
4	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS	Pb	pH	T			
5	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS	Pb	pH				
6	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS	Pb					
7	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS	TSS						
8	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD	TDS							
9	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl	BOD								
10	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr	Cl									
11	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd	Cr										
12	COD	TOC	OG	NH_3N	As	Ni	Zn	Cd											
13	COD	TOC	OG	NH_3N	As	Ni	Zn												
14	COD	TOC	OG	NH_3N	As	Ni													
15	COD	TOC	OG	NH_3N	As														
16	COD	TOC	OG	NH_3N															
17	COD	TOC	OG																
18	COD	TOC																	
19	COD																		

of each factor's squares equals the index's square (Eq. (4)). The WQI is conceptualized in this approach as a three-dimensional space, with each axis representing one of the three factors. According to the CCME's definition, the index is directly related to these factors.

$$WQI = \left[\frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732}\right] \quad (4)$$

The calculated values are normalized using the divisor 1.732, so the Water Quality Index (WQI) result is from 0 to 100. On this scale, the water quality is evaluated from 0 to 100, with 100 representing the highest quality. The WQI was calculated using the CCME Calculator Software Version 1.0 given by Canadian Council of Ministers of the Environment. The CCME has established five categories based on water quality can be seen in Table 1.

2.4. Machine-Learning Models

Six machine-learning models (AdB, K-NN, GB, RF, SVR and BR) were chosen to be trained on the water quality dataset. In making this choice, their performance indicators were considered, and an assessment was conducted regarding how these models are aligned with specific requirements (Hillel et al., 2021). AdB model was selected due to its expertise in handling complex datasets and excellent accuracy. A strong learner is created using the ensemble method by combining an array of weak learners. The overall predictive power is increased by the boosting method used in AdB (Bourel and Segura, 2018; Tanha et al., 2020). However, parameter tuning may be necessary to improve performance because it can be sensitive to noisy data and outliers. K-NN performs well when the underlying data distribution is locally homogeneous and displays nonlinear patterns because it captures complex relationships in the data because of its adaptability and simplicity (Chen et al., 2020). However, K-NN requires careful consideration when choosing the number of neighbors (K) and an appropriate distance metric, especially for large datasets (Ahmed et al., 2019). Gradient Boost was a good option because it can effectively capture intricate interactions in the data (Zhou et al., 2020). It is an ensemble method that builds models one at a time, correcting the flaws of the previous model as it goes. GB is skilled at handling heterogeneous data and generating highly accurate predictions (Polikar, 2012). However, if the learning rate is slow or the number of boosting iterations is excessive, it may be subject to overfitting. Hyperparameters need to be tuned carefully (Uddin et al., 2022). RF can manage interactions, outliers, and high-dimensional data. This collection of decision trees lessens overfitting via averaging predictions from various trees (Wang et al., 2021). The robustness, computational effectiveness, and provision of feature importance measures of RF are well known. In contrast to other ensemble methods, however, RF can be more challenging to interpret than a single decision tree and may have trouble capturing complex nonlinear relationships (Rigatti, 2017). SVR shows excellent potential when datasets have numerous dimensions, complex relationships, and nonlinear patterns (Ahmed et al., 2019). It is less likely to over fit and is better able to handle outliers. SVR offers a range of kernel functions to capture various data patterns. SVR can be computationally expensive for large datasets and requires careful preprocessing and data scaling (Modaresi and Araghinejad, 2014). Choosing the appropriate kernel functions and tuning the hyperparameters can be challenging. The probabilistic framework of BR, along with its aptitude for handling uncertainty estimation and model choice, led to its selection. It is resistant to overfitting and capable of handling intricate data relationships (Tanha et al., 2020). However, BR might have trouble capturing intricate nonlinear patterns compared to other models. It presupposes a particular type of prior distribution, which might not always correspond to the underlying data distribution (Holmgren et al., 2014).

2.5. Research methodology

Our approach involved six ML models, including Adaptive Boosting, K-Nearest Neighbors, Gradient Boosting, Random Forests, Support Vector Regression, and Bayesian Regression to predict the Water Quality Index (WQI) of the Aik-Stream based on 150 samples and 19 parameters. We randomly split the dataset into training (80%) and testing (20%) subsets to address the issue of overfitting in the model. The training subset helped us in understanding the system's behavior, while the testing subset was used for model validation (Nguyen et al., 2021). We also employed the feature elimination-linear algorithm RFE-L to select relevant features from nineteen possible input variable combinations. This process helped in streamlining the model by reducing the number of variables (Ebrahimi-Khusfi et al., 2021). Each model was trained with nineteen different input combinations, labeled 1 to 19 (as detailed in Table 2).

Our methodology follows the "Wrapper" approach as described by (Kohavi and John, 1997) which involves a predetermined selection of input combinations for efficiency and robustness. Normalization of all input water quality parameters between 0.1 and 0.9 was performed to ensure uniformity in variable values, preventing bias towards specific ranges. A top-down sequential search algorithm was used for model evaluation, where a variable is removed at each step to see if the model's precision is maintained. This process continues until no more variables can be eliminated without dropping below the acceptable precision

Table 3
Descriptive statistics of water quality parameters of the Aik-Stream, including sample size [N = 150 (3x)].

Parameters	Mean			Standard deviation			Kurtosis			Skewness			Permissible Limit
	LPZ	HPZ	MPZ	LPZ	HPZ	MPZ	LPZ	HPZ	MPZ	LPZ	HPZ	MPZ	WHO (mg/L)
pH	7.49	6.18	7.94	0.78	0.767	1.295	−0.09	−1.21	−0.002	−0.32	−0.31	−1.09	6.5–7
COD mg/L	6.054	62.39	43.15	6.05	16.972	15.62	−0.62	0.65	1.30	1.52	0.02	1.46	250
BOD mg/L	12.02	43.02	24.56	12.0	6.733	9.55	−0.34	−0.05	0.36	−0.48	−0.46	−0.55	50
TDS mg/L	6.24	7.00	6.87	0.42	0.000	0.33	1.28	0.00	−2.34	0.36	0.00	3.63	300–600
TSS mg/L	5.94	6.00	0.24	0.00	0.000	0.46	−0.72	0.00	0.07	−0.47	0.00	−0.39	250
NH_3–N mg/L	5.96	22.73	4.424	2.56	3.447	4.42	0.07	−0.64	−0.83	−1.53	−0.50	−1.36	50
O&G mg/L	2.12	5.75	2.19	0.81	1.309	1.06	−0.22	−0.50	0.60	−1.46	−0.24	−0.36	10
TOC mg/L	13.6	110.9	40.33	4.55	17.512	28.12	−0.54	−0.28	0.65	0.91	−0.96	−1.31	15
NO_3–N mg/L	0.45	0.51	0.17	0.50	0.505	0.37	0.20	−0.04	1.84	−2.04	−2.08	1.47	15
Cl^- mg/L	157.4	221.2	151.6	44.2	28.915	45.15	−0.12	−0.30	0.46	−0.72	−0.48	−0.69	200
Cu mg/L	0.47	0.50	0.58	0.50	0.140	0.49	0.121	−7.14	−0.34	−2.06	51.0	−1.96	1
Zn mg/L	−0.14	1.00	0.23	0.34	0.316	0.42	−2.173	−0.46	1.33	2.83	8.02	−0.24	3
Cr mg/L	0.20	1.94	1.27	0.40	0.947	0.64	1.578	0.70	2.68	0.50	−0.41	7.39	0.05
Pb mg/L	0.76	2.12	1.44	0.42	0.739	0.58	−1.286	0.11	0.93	−0.36	−0.41	−0.06	0.05
Cd mg/L	0.00	0.98	0.40	0.00	0.510	0.494	0.00	−0.03	0.44	0.00	1.14	−1.88	0.05
Ni mg/L	0.00	0.90	0.54	0.00	0.300	0.50	0.00	−2.78	−0.17	0.00	5.99	−2.05	0.1
As mg/L	0.00	1.53	0.83	0.00	0.504	0.72	0.00	−0.12	0.26	0.00	−2.06	−1.01	0.05
Hg mg/L	0.00	0.20	0.06	0.00	0.401	0.24	0.00	1.57	3.73	0.00	0.50	12.44	0.02

level. This approach, include the feature selection method, aligns with prior research in the field (Mehdizadeh et al., 2020; Mokhtar et al., 2021).

2.6. Evaluation metrics

We conducted a comparison between the observed WQI, and the values predicted by our studied models. The precision and accuracy of the models was quantified using different statistical metrics, namely Mean Absolute Error (MAE) (Eq. (5)), Root Relative Squared Error (RRSE) (Eq. (6)), Root Mean Square Error (RMSE) (Eq. (7)), and the coefficient of determination (R^2) (Eq. (7)). These specific metrics were chosen based on the recommendations of previous studies (Malone et al., 2017). The parameters are defined as: "WQI actual" represents the observed or actual WQI value, while "WQI predicted" represents the simulated or predicted WQI value.

$$\mathrm{MAE} = \left(\frac{1}{n}\right) * \Sigma\left|\mathrm{WQI_{pred}} - \mathrm{WQI_{actual}}\right| \tag{5}$$

$$\mathrm{RRSE} = \left(\frac{1}{n}\right) * \Sigma(\mathrm{WQI_i^P} - \mathrm{WQI_i^A})^2 \tag{6}$$

$$\mathrm{RMSE} = \mathrm{sqrt}\left(\left(\frac{1}{n}\right) * \Sigma\left(\mathrm{WQI_{pred}} - \mathrm{WQI_{actual}}\right)^2\right) \tag{7}$$

$$R^2 = 1 - \left(\frac{\Sigma\left(\mathrm{WQI_{actual}} - \mathrm{WQI_{pred}}\right)^2}{\Sigma(\mathrm{WQI_{actual}} - \mathrm{mean}(\mathrm{WQI_{actual}}))^2}\right) \tag{8}$$

A higher R-squared value indicates a more robust correlation or fit between the observed and actual values (Zamani et al., 2023). On the other hand, lower values of MAE, MSE, and RMSE indicate improved model performance (Chicco et al., 2021). These evaluation metrics were employed to evaluate how well the regression models predicted the WQI.

2.7. Feature selection

Feature selection is defined as a search process that seeks to extract a relevant subset of attributes from the original collection. There are other ways to choose features; in this work, we use the Recursive Feature Elimination-Linear (RFE-L) technique. Since RFE-L uses a backward selection method to find the best feature combination for predicting the target variable, it is a popular approach for finding the most relevant features in predictive modeling (Akhtar et al., 2020; Ebrahimi-Khusfi et al., 2021). Initially, the algorithm constructs a model using all available features and computes the importance of each feature within the model. Subsequently, it ranks these features and systematically eliminates the least significant ones based on the model's evaluation metrics, such as RMSE and R^2 (Bagherzadeh et al., 2021). The model is then retrained, and the importance of the independent variables is reassessed. This iterative process continues until a specific number of predictive subsets are identified, enabling the assessment or selection of the subset of predictor variables, which are the water quality parameters in this context. The size of the subset is determined to choose the most optimal predictor variables (Kuhn and Johnson, 2018). It's important to note that in this algorithm, the ideal combination of features is achieved when the values of RMSE approach 0 and those of R^2 approach 1. This signifies the best model fit for the given dataset.

2.8. Feature importance score

Machine learning models evaluate input variables or features by assigning them importance scores or weights based on their internal computations (Siham et al., 2021). These scores represent the extent to which each input variable impacts the model's predictions. When comparing various machine learning models, you can evaluate how they diverge in their allocation of importance scores. Some models may prioritize specific features, while others may emphasize different ones (Singh et al., 2022). This insight can inform your selection of the most appropriate model for a given task. Moreover, the choice of input features plays a pivotal role in the stability and resilience of the prediction model. By thoughtfully examining and choosing relevant input variables, you can enhance the models' stability, ensuring consistent and dependable predictions across different scenarios (Gültekin and Sakar, 2018; Hameed et al., 2017; Singha et al., 2021; Wong et al., 2022). The relative importance score of each input feature is determined by the optimization algorithm used by the respective prediction models. Different models assign varying levels of importance score to the input variables when predicting the target variable (Gültekin and Sakar, 2018; Singha et al., 2021).

3. Results

3.1. Water Physicochemical Properties

The results showed a slight difference in the pH of two consecutive sampling sites, ranging from 6.15 to 8.95 (Table 3). The temperature stayed constant around 28.0 °C. This can be attributed to the buffering

Table 4
Pearson correlation matrix of all studied variables.

Variables	COD	BOD	TDS	pH	T	NH_3N	TSS	OG	TOC	NO_3N	Cl	Cu	Zn	Cr	Pb	Cd	Ni	As	Hg
COD	1																		
BOD	0.69**	1																	
TDS	0.84**	0.57*	1																
pH	−0.57*	−0.56*	−0.6*	1															
T	0.19	0.42**	−0.03	−0.1	1														
NH_3–N	0.82**	0.83**	0.74**	−0.64*	0.43*	1													
TSS	0.65**	0.63*	0.59*	−0.52*	0.24	0.71**	1												
OG	0.82**	0.75**	0.80**	−0.69*	0.17	0.83**	0.79*	1											
TOC	0.76**	0.82**	0.72**	−0.73**	0.28	0.87**	0.86**	0.87**	1										
NO_3N	0.02	−0.12	0.13	0.12	−0.16	−0.12	0.09	0.07	−0.04	1									
Cl	0.63**	0.69**	0.55**	−0.60*	0.38*	0.76**	0.83**	0.82**	0.85**	0.16	1								
Cu	0.67**	0.71*	0.61**	−0.55*	0.31*	0.78**	0.66*	0.72*	0.76*	−0.15	0.61*	1							
Zn	0.68**	0.79*	0.68**	−0.58*	0.21	0.76**	0.55*	0.70*	0.78**	0.03	0.67*	0.61*	1						
Cr	0.60**	0.81**	0.53*	−0.57*	0.49*	0.84**	0.53**	0.69*	0.76**	−0.07	0.72*	0.66*	0.71*	1					
Pb	0.61**	0.68*	0.58*	−0.49*	0.33*	0.76**	0.49*	0.78**	0.66*	0.02	0.60*	0.69*	0.62**	0.74*	1				
Cd	0.59**	0.76**	0.55*	−0.54*	0.48*	0.82**	0.48*	0.71*	0.69*	−0.02	0.68*	0.59*	0.70*	0.85*	0.76*	1			
Ni	0.61**	0.73*	0.51**	−0.59*	0.58**	0.79**	0.69*	0.71*	0.79**	−0.05	0.71*	0.66*	0.66*	0.71*	0.69*	0.69*	1		
As	0.74**	0.84**	0.73*	−0.70**	0.28	0.88**	0.71**	0.86**	0.90*	0.02	0.82**	0.73**	0.86**	0.83**	0.74*	0.80**	0.72**	1	
Hg	0.42**	0.42*	0.42*	−0.45*	0.27	0.55*	0.57*	0.57*	0.56*	−0.07	0.50*	0.61*	0.36*	0.42*	0.54*	0.45*	0.62*	0.52*	1

**Correlation coefficient significant p value at 0.01; * Correlation coefficient p value significant at p value 0.05.

capacity of water playing a role in maintaining pH levels and the thermal inertia of the water body helps to slow down temperature fluctuations. The amount of TDS, excessive values from 439 mg/L to 1340 mg/L were observed, and in TSS, from 201 to 334 mg/L. The various forms of nitrogen content closely aligned the patterns observed in BOD and COD, indicating that elevated nitrate-nitrogen (NO_3–N) and ammonia nitrogen (NH_3–N) concentrations are associated with increased pollutant loads. The total organic carbon content oscillated from 0 to 90 mg/L. The minimum value of chloride recorded was 74.22 mg/L. Noticeable differences in the measurements of heavy metals were observed if moved from the upstream region to the midstream areas. This variation has divided the stream into three distinct zones. These zones were named based upon the special differences in water quality as Less Polluted Zone (LPZ) for the upstream area based on its fair and reasonable water quality. In contrast, the midstream area with poor water quality and high pollution load is represented as a Highly Polluted Zone (HPZ), and the area of a downstream area with somewhat less reduced water quality is defined as a Moderately Polluted Zone (MPZ). The High Polluted Zone (HPZ) represents areas with significant water contamination due to point sources like tanneries, industrial effluents, and municipal sewage from Sialkot City, leading to highly degraded water quality. In contrast, Moderate Polluted Zone (MPZ) includes sites further downstream, affected by non-point sources and domestic sewage from nearby smaller towns. As the water travels, sedimentation occurs, which aids in reducing the pollution levels. Furthermore, when this water merges with the Palkhu stream at various confluence points downstream, its quality improves due to dilution. This natural dilution

Table 5
Evaluation measures for the GB algorithm in WQI prediction for the aik-stream (training and testing datasets).

Model	R^2	RMSE	MAE	RAE (%)	RRSE (%)
Training					
1	0.93	6.45	9.35	82.23	21.57
2	0.79	7.32	5.37	99.45	53.22
3	0.75	4.38	10.81	109.11	62.82
4	**0.94**	**6.33**	**4.62**	**38.21**	**39.71**
5	0.82	6.47	11.12	70.55	58.67
6	0.75	7.37	6.25	91.32	50.09
7	0.75	4.46	7.35	53.25	87.91
8	0.73	7.56	11.45	44.45	51.81
9	0.83	9.56	13.21	35.65	89.22
10	0.81	8.26	12.23	62.12	91.45
11	0.80	8.67	12.56	110.23	83.21
12	0.77	6.61	7.41	37.23	72.36
13	**0.88**	**7.24**	**10.15**	**16.34**	**73.56**
14	0.72	5.23	5.85	99.23	83.56
15	0.83	4.45	7.57	98.45	64.34
16	0.87	4.2	12.32	71.67	104.67
17	0.71	7.34	4.23	98.89	96.67
18	0.78	8.35	7.23	76.34	67.34
19	0.85	7.36	4.23	101.23	101.56
Testing					
1	0.91	7.05	9.15	87.21	27.01
2	0.84	11.34	13.63	83.23	118.56
3	0.77	10.76	8.53	48.54	85.32
4	**0.90**	**6.54**	**8.65**	**59.32**	**55.32**
5	0.79	11.43	7.54	50.23	77.32
6	0.66	10.32	13.76	37.31	86.32
7	0.74	8.31	12.98	38.41	65.32
8	0.70	8.54	11.76	90.12	60.34
9	0.81	7.43	12.32	48.23	86.12
10	0.77	10.25	13.32	57.67	53.23
11	0.79	10.21	7.31	59.34	52.45
12	0.72	11.24	12.41	38.65	72.67
13	**0.85**	**8.67**	**11.31**	**19.31**	**8.32**
14	0.84	7.54	8.31	42.45	110.56
15	0.85	10.25	8.21	118.12	113.78
16	0.64	12.43	7.41	119.23	112.54
17	0.72	8.67	9.32	81.21	85.23
18	0.75	6.78	9.31	58.12	57.3
19	0.73	8.67	11.31	41.12	109.23

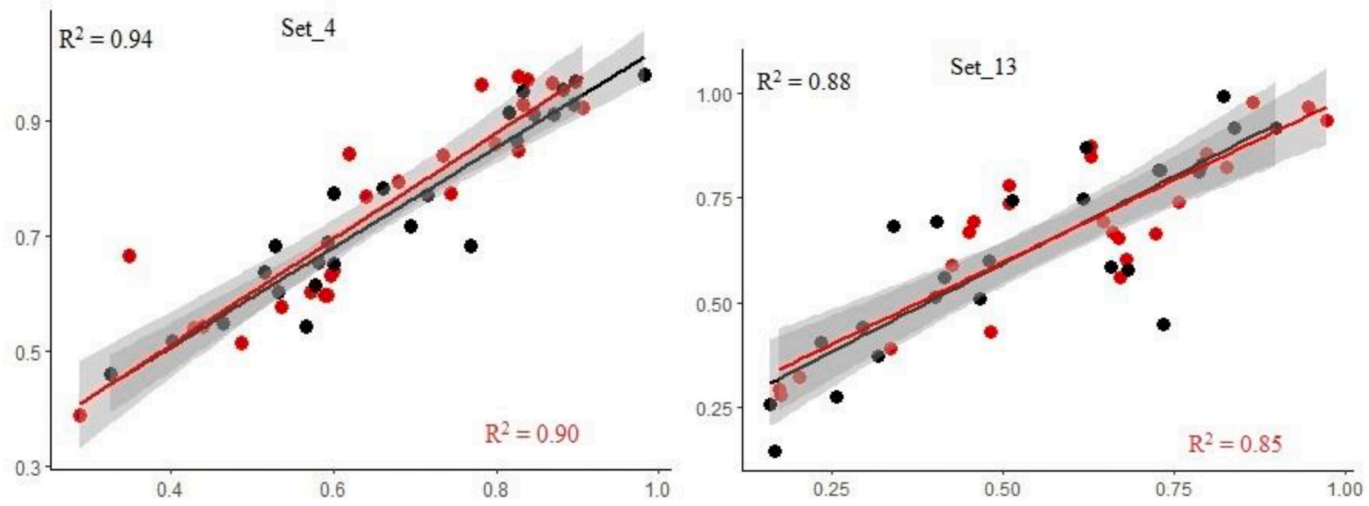

Fig. 3. a–b.GB Algorithm's Best Inputs (4th and 13th) for WQI Prediction (Training: Black, Testing: Red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

process plays a crucial role in gradually changing water quality in the region.

Pearson Correlation matrix highlights significant associations between physicochemical parameters and heavy metals and metalliods. There are noteworthy positive correlations between Chemical Oxygen Demand (COD) and Biological Oxygen Demand (BOD) with ammonia nitrogen (NH3–N). Similarly, the matrix reveals strong correlations among heavy metals themselves. For example, chromium and cadmium have a strong positive correlation (r = 0.85). Likewise, arsenic shows significant correlations with zinc (r = 0.87), cadmium (r = 0.81), and chromium (r = 0.84). (Table 4).

3.2. Calculation of WQI

Water samples from the Aik-Stream are categorized into three classes based on the water quality index (WQI) as fair, marginal, and poor. This classification provides insights into the varying water quality levels observed in the stream (Supplementary Table 2). The observed WQI demonstrated that the upstream portion with 51 monitoring sites was of good and fair quality (64 < WQI ≤ 94), while the midstream with 66 monitoring sites mainly was classified as poor water quality (0 < WQI ≤ 44). The downstream portion with 33 monitoring sites had marginal values (45 < WQI ≤ 64). The water quality in the upstream region is comparatively better and exhibits minimal contamination, primarily due to the absence of industrialization (NN & NN, pers. obs.). Nevertheless, once it traverses through the city, industrial and sewage waste pollutants gradually accumulate, resulting in a substantial build-up of various contaminants, including heavy metals. Thus, the water at the mid-stream undergoes excessive pollution, threatening its overall quality. Moving downstream, the water becomes moderately polluted due to a decline in industrial activity in that area. The deterioration of water quality from upstream to downstream of Aik-Stream is mainly linked to discharges of urban and industrial wastewater in that area.

3.3. Evaluating ML models

A set of statistical metrics, including R^2, RMSE, MAE, RAE and RRSE, were employed to evaluate the predictive performace of the WQI during both the training and testing phases (Tables 5–10).

The evaluation of the GB model's predictive performance, both during training and testing phases, shown that among the various GB models investigated in this study, the GB-4th input combination (with training R^2 = 0.94, training RMSE = 6.33, testing R^2 = 0.90, testing RMSE = 6.54) demonstrated the most optimal performance (Table 5). It is important to note that predictive performance typically leans more favorably towards the training phase as compared to the testing phase, which is an expected outcome during the model training process aimed at minimizing predictive errors. The GB model is initially constructed

Table 6
Evaluation measures for the RF algorithm in WQI prediction for the aik-stream (training and testing datasets).

Model	R^2	RMSE	MAE	RAE (%)	RRSE (%)
Training					
1	0.89	7.45	10.34	89.45	99.76
2	**0.93**	**5.23**	**6.45**	**32.84**	**55.33**
3	0.78	8.45	11.56	75.54	64.32
4	0.63	8.45	14.67	25.23	31.76
5	0.71	8.45	14.56	52.43	72.23
6	0.73	13.78	10.81	96.45	94.24
7	0.87	6.9	10.89	70.56	60.62
8	0.61	12.27	13.32	92.23	85.23
9	0.78	8.67	13.76	108.34	103.98
10	**0.85**	**8.67**	**9.23**	**35.21**	**72.34**
11	0.75	12.67	6.57	56.56	75.12
12	0.79	8.78	12.87	18.56	83.23
13	0.88	9.34	14.52	86.56	80.34
14	0.70	5.67	11.51	113.45	65.45
15	0.77	6.98	12.65	86.78	116.23
16	0.79	8.45	12.54	66.56	65.56
17	0.67	13.67	8.67	74.32	79.12
18	0.89	7.67	14.23	79.45	74.23
19	0.87	7.98	12.26	88.12	60.09
Testing					
1	0.75	10.92	7.12	96.41	81.21
2	**0.91**	**8.02**	**9.24**	**51.34**	**75.34**
3	0.74	7.23	10.91	97.21	51.45
4	0.82	11.34	10.98	48.19	86.45
5	0.77	12.61	8.56	80.81	78.87
6	0.83	11.34	6.54	64.72	67.34
7	0.84	9.45	12.45	98.63	64.32
8	0.71	10.45	11.63	62.53	67.35
9	0.75	10.45	12.45	93.49	74.45
10	**0.84**	**9.34**	**10.76**	**58.61**	**64.23**
11	0.73	9.45	8.47	85.54	85.45
12	0.74	9.45	13.87	71.51	70.45
13	0.76	8.87	12.65	78.34	81.56
14	0.65	9.65	11.73	121.61	58.31
15	0.75	10.65	6.63	81.61	83.75
16	0.75	10.43	11.52	92.71	60.56
17	0.64	5.89	12.61	84.87	88.23
18	0.84	7.62	7.52	64.23	76.67
19	0.67	6.71	10.76	48	85.23

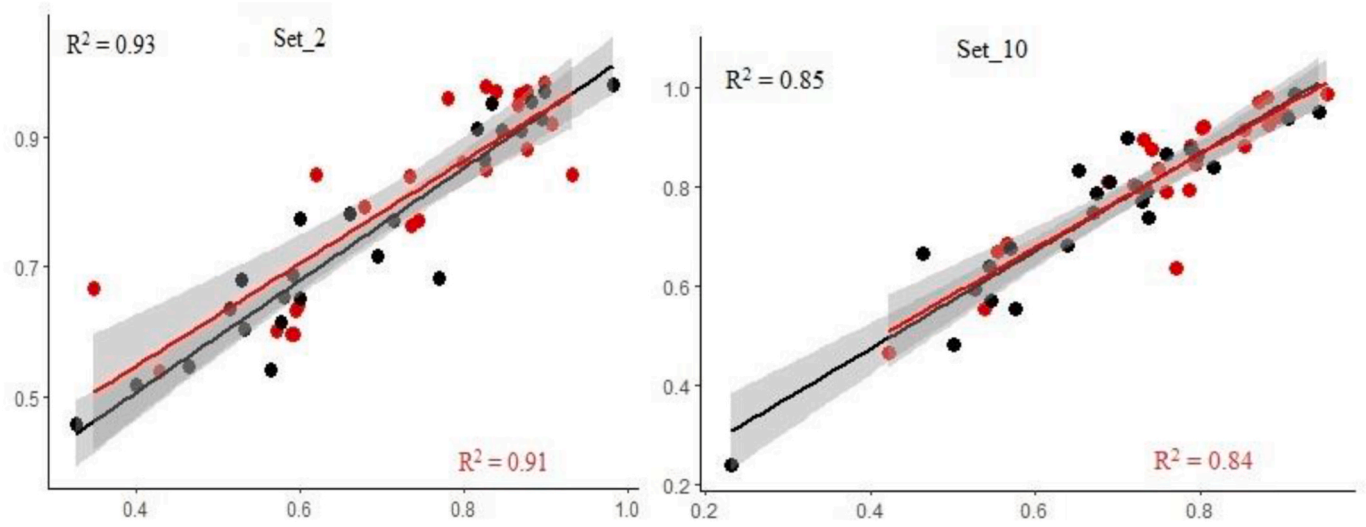

Fig. 4. a–b RF Algorithm's Best Inputs (2nd and 10th) for WQI Prediction (Training: Black, Testing: Red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 7
Evaluation measures for the BR algorithm in WQI prediction for the aik-stream (training and testing datasets).

Model	R^2	RMSE	MAE	RAE (%)	RRSE (%)
Training					
1	0.91	10.12	10.98	91	81.67
2	0.92	6.34	10.76	94	99.76
3	0.84	12.87	7.56	117	70.65
4	0.79	11.87	12.54	117	79.65
5	0.83	12.76	7.76	78	95.65
6	**0.93**	**5.65**	**6.65**	**120**	**83.54**
7	0.78	11.45	13.76	102	79.43
8	0.76	7.65	9.34	117	131.43
9	0.71	9.54	8.45	126	128.67
10	0.76	6.78	13.56	80	118.98
11	0.83	9.67	7.56	124	117.43
12	0.73	11.65	10.68	87	89.34
13	**0.87**	**7.34**	**13.78**	**64**	**88.76**
14	0.85	10.34	12.45	76	135.87
15	0.79	6.34	8.34	117	120.74
16	0.73	12.67	12.21	80	107.43
17	0.82	7.57	8.45	74	120.45
18	0.73	7.67	10.56	130	88.76
19	0.81	10.45	11.45	75	84.54
Testing					
1	0.86	7.98	10.83	110.87	103.34
2	0.85	13.23	9.72	92.23	83.23
3	0.83	10.23	11.45	114.45	77.45
4	0.68	9.56	10.72	104.67	74.34
5	0.73	8.45	8.91	92.87	84.45
6	**0.89**	**8.45**	**14.56**	**116.34**	**91.65**
7	0.61	14.45	14.34	123.34	115.23
8	0.68	14.67	11.67	136.34	121.87
9	0.84	9.78	13.93	90.12	133.23
10	0.75	10.54	14.73	118.14	106.12
11	0.69	7.72	10.52	108.23	93.56
12	0.87	11.34	10.95	103.56	137.89
13	**0.84**	**10.47**	**10.45**	**83.67**	**91.34**
14	0.87	7.12	9.67	109.75	100.12
15	0.72	11.23	11.93	128.14	122.67
16	0.73	14.62	14.34	116.56	94.34
17	0.85	12.56	12.67	88.45	79.98
18	0.73	11.56	12.34	128.76	88.45
19	0.81	8.67	14.56	135.23	101.23

during the training stage and subsequently assessed during the testing phase. Furthermore, the GB–13th input combination also exhibited better predictive performance in both the training (R^2 = 0.88, RMSE = 7.24) and testing (R^2 = 0.85, RMSE = 8.67) stages when compared to other input combination models. Although the GB–4th input combination model, consisting of sixteen input parameters (predictors), offers the best predictive performance, the GB–13th input combination model, which incorporates only seven input parameters, namely COD, TOC, OG, NH_3N, As, Ni, and Zn, also provides satisfactory performance (R^2 train = 0.88, R^2 test = 0.85). Therefore, the GB–13th input combination model is identified as the superior predictive model, requiring a more limited number of input physicochemical variables. Overall, the results demonstrate that the GB model delivers a high level of predictive accuracy for water quality in both the training and testing stages (Table 5). The visual representation of the GB's best input combinations can be seen in Fig. 3a–b as training (black) testing (red).

The RF model's predictive performance is notable. The RF-2nd input combination demonstrates exceptional predictive capabilities in both the training and testing phases (R^2 = 0.93, RMSE = 6.33) and the testing phase (R^2 = 0.85, RMSE = 8.02) (Table 6 and Fig. 4a–b). However, the 10th input combination, despite utilizing a smaller number of variables, also delivers commendable results in both training (R^2 = 0.85, RMSE = 8.67) and testing (R^2 = 0.84, RMSE = 9.34).

The performance of the BR predictive models, highlighting the superior performance of the BR-6th and 13th input combinations. These combinations demonstrate remarkable proficiency, with the BR-6th combination achieving high scores in both the training phase (R^2 = 0.93, RMSE = 5.65) and the testing phase (R^2 = 0.89, RMSE = 8.45), while the BR-13th combination also exhibits robust results in training (R^2 = 0.87, RMSE = 7.34) and testing (R^2 = 0.84, RMSE = 10.47) (Table 7 and Fig. 5a–b). The Ada Boost model's performance. The 7th and 10th input combination stands out as a top performer (training R^2 = 0.90, training RMSE = 7.91, testing R^2 = 0.89, testing RMSE = 8.41) and (training R^2 = 0.89, training RMSE = 10.22, testing R^2 = 0.87, testing RMSE = 10.45) respectively (Table 8 and Fig. 6a–b).

KNN model shows that the 2nd input combination performed well in both training (R^2 = 0.88, RMSE = 8.32) and testing (R^2 = 0.83, RMSE = 11.34) (Table 9 and Fig. 7a–b). The 7th input combination also had good results with R^2 values of 0.87 in training and 0.84 in testing. However, when assessing the predictive performance of the SVM model, it becomes evident that the SVM-6th input combination stands out as a robust performer, delivering superior results in both the training (R^2 = 0.85, RMSE = 5.41) and the testing phase (R^2 = 0.82, RMSE = 6.67) than other input combinations (Table 10 and Fig. 8a–b).

The additional scatter plots depicting all input combinations for comprehension in each model can be found in the supplementary data section for further reference (Supplementary Figures 1- 6).

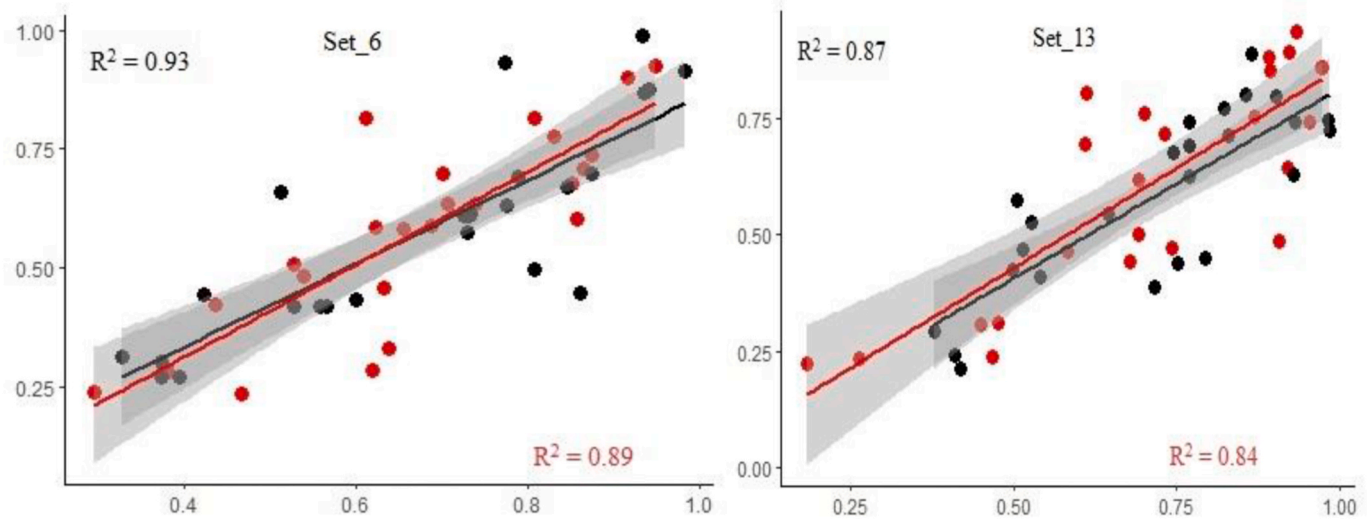

Fig. 5. a–b BR Algorithm's Best Inputs (6th and 13th) for WQI Prediction (Training: Black, Testing: Red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 8
Evaluation measures for the AdB algorithm in WQI prediction for the aik-stream (training and testing datasets).

Model	R^2	RMSE	MAE	RAE (%)	RRSE (%)
Training					
1	0.76	8.35	12.23	89.52	105.23
2	0.78	9.65	8.24	115.21	123.54
3	0.82	11.23	9.82	64.31	68.98
4	0.72	5.98	10.61	113.34	75.63
5	0.69	10.72	9.52	85.92	105.35
6	0.54	5.34	9.73	116.87	115.23
7	**0.90**	**7.91**	**8.23**	**93.43**	**84.56**
8	0.88	5.99	7.34	61.34	101.98
9	0.72	8.001	13.34	71.34	115.23
10	**0.89**	**10.22**	**13.62**	**72.56**	**89.34**
11	0.65	12.87	10.87	60.87	107.45
12	0.64	5.92	9.65	125.43	83.72
13	0.64	6.23	8.53	98.71	89.56
14	0.53	8.54	11.62	126.23	113.98
15	0.85	7.52	9.62	62.62	88.41
16	0.79	11.23	8.9	125.45	67.34
17	0.53	11.23	11.09	128.67	83.43
18	0.84	5.23	11.52	90.23	109.34
19	0.81	11.24	10.34	120.23	117.23
Testing					
1	0.81	7.35	8.15	131.21	127.01
2	0.80	11.44	10.63	116.23	128.56
3	0.79	10.76	7.53	58.54	95.32
4	0.69	6.84	5.65	69.32	75.32
5	0.75	11.93	9.54	121.23	117.32
6	0.42	10.72	10.76	38.31	126.32
7	**0.89**	**8.41**	**12.98**	**39.41**	**85.32**
8	0.77	8.64	10.76	101.12	111.34
9	0.70	7.73	10.32	49.23	121.12
10	**0.87**	**10.45**	**11.32**	**87.67**	**93.23**
11	0.82	10.71	8.31	69.34	92.45
12	0.75	11.94	13.41	48.65	73.67
13	0.57	8.77	12.31	113.31	193.32
14	0.88	7.74	9.31	45.45	110.56
15	0.81	10.65	7.21	108.12	123.78
16	0.63	12.73	9.41	109.23	112.54
17	0.36	8.47	8.32	84.21	95.23
18	0.77	6.58	6.31	55.12	97.3
19	0.75	8.67	10.31	48.12	119.23

The overall comparison analyses of the ML models showed that the GB, RF, BR and AdB models demonstrated robust performance, whereas the SVR and K-NN models displayed average performances in accurately modeling the water quality index. In particular, the Gradient Boost model distinguished itself from other models with superior predictive abilities and its (GB–13th) input combination model is depicted to be the better predictive model, which needs a limited number of input physicochemical variables.

3.4. Identifying optimal input combinations

We also carried out a best-subset regression analysis to identify the optimal input combinations for our Water Quality (WQ) model. To achieve this, we computed six statistical criteria, including Mean Squared Error (MSE), determination coefficients (R^2), adjusted R^2, Mallows' Cp (Gilmour, 1996), Akaike's AIC, and BIC. The results of these computations clearly indicate that Model 15 (comprising TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N, and Zn) emerges as the favored choice among the diverse models. It displayed the lowest Mean Squared Error (MSE) at 4.034, the lowest Akaike Information Criterion (AIC) at 351.54, the lowest Bayesian Information Criterion (BIC) at 579.42, the lowest Mallows' Cp value at 10.54, as well as the highest R^2 value at 0.93 and adjusted R^2 at 0.92. As a result, Model 15 was identified as the most suitable input combination for predicting the WQI model (Table 11).

Moreover, to assess the potential presence of multicollinearity among the water quality index parameters, we utilized the Variance Inflation Factor (VIF) and its reciprocal, 1/VIF, as presented in Table 12. A widely accepted criterion is that 1/VIF should be less than 0.1, and VIF should be less than 10 to indicate the absence of multicollinearity in relation to the target variable, WQI. The findings in Table 12 reveal that the majority of variables met the criteria with 1/VIF values below 0.1. However, during the VIF assessment, NH_3N exhibited a value of 10.32, and As had a value of 11.43, both slightly surpassing the VIF threshold of 10. Nevertheless, these results do not suggest the presence of multicollinearity.

3.5. Feature importance and ML models

The input features significantly influence the stability and robustness of the prediction model. Stability of the models can be improved, guaranteeing consistent and reliable predictions across various states by carefully considering and selecting the appropriate input variables (Gültekin and Sakar, 2018; Singha et al., 2021). The relative importance of input features is determined by the optimization algorithm used by the respective prediction models. Different algorithms may assign

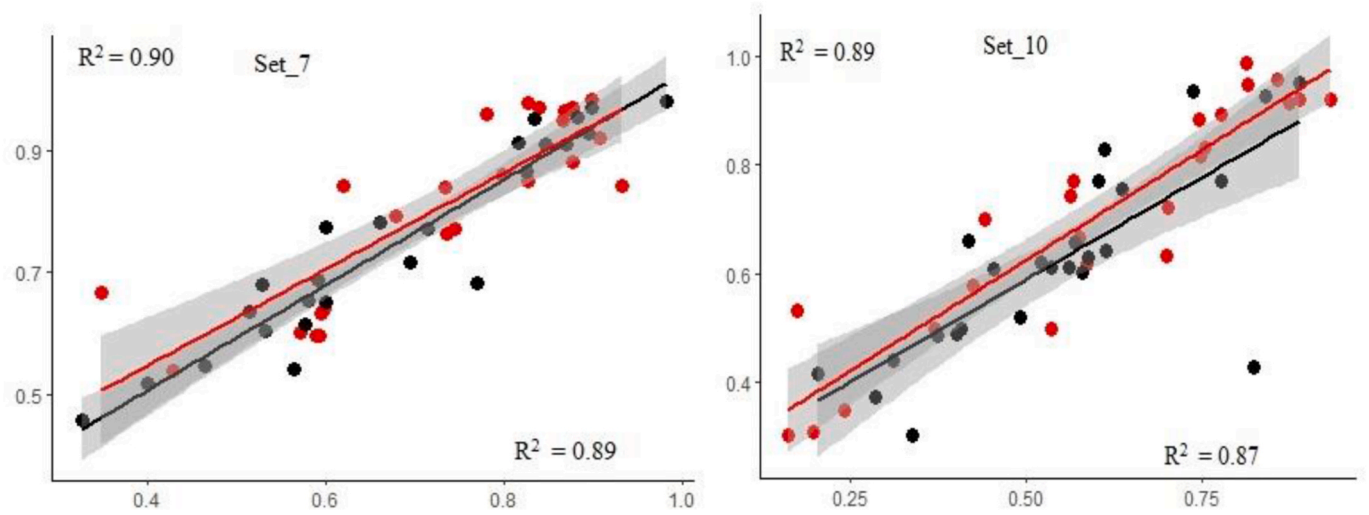

Fig. 6. a–b AdB Algorithm's Best Inputs (7th and 10th) for WQI Prediction (Training: Black, Testing: Red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 9
Evaluation measures for the K-NN algorithm in WQI prediction for the aikstream (training and testing datasets).

Model	R^2	RMSE	MAE	RAE (%)	RRSE (%)
Training					
1	0.87	6.45	8.15	83.23	91.57
2	**0.88**	**8.32**	**10.63**	**99.45**	**53.22**
3	0.78	4.38	17.53	109.11	62.82
4	0.63	8.73	5.65	108.21	39.71
5	0.48	6.47	9.54	70.55	108.67
6	0.43	7.37	10.76	91.32	50.09
7	**0.87**	**4.46**	**6.98**	**53.25**	**37.91**
8	0.81	8.56	10.76	44.45	51.81
9	0.78	9.56	10.32	35.65	89.22
10	0.84	8.26	11.32	63.12	91.45
11	0.63	8.67	8.31	110.23	83.21
12	0.53	6.61	13.41	37.23	72.36
13	0.58	7.24	12.31	96.34	73.56
14	0.82	8.23	19.31	99.23	83.56
15	0.77	4.45	7.21	98.45	64.34
16	0.50	8.2	9.41	71.67	104.67
17	0.52	7.34	8.32	98.89	96.67
18	0.79	8.35	6.31	76.34	67.34
19	0.77	7.36	10.31	101.23	101.56
Testing					
1	0.80	8.05	9.15	101.21	117.01
2	0.83	11.34	13.63	116.23	128.56
3	0.79	10.76	8.53	48.54	85.32
4	0.65	6.54	8.65	59.32	55.32
5	0.78	11.43	7.54	121.23	137.32
6	0.46	10.32	13.76	37.31	116.32
7	**0.84**	**9.31**	**12.98**	**38.41**	**65.32**
8	0.35	8.54	11.76	101.12	130.34
9	0.51	7.43	12.32	48.23	121.12
10	0.46	10.25	13.32	57.67	83.23
11	0.76	10.21	7.31	59.34	82.45
12	0.72	11.24	12.41	38.65	72.67
13	0.53	8.67	11.31	103.31	93.32
14	0.74	8.54	8.31	42.45	110.56
15	0.75	10.25	9.21	108.12	133.78
16	0.64	12.43	7.41	109.23	152.54
17	0.82	8.67	9.32	81.21	185.23
18	0.45	6.78	9.31	59.12	97.3
19	0.73	8.67	11.31	41.12	119.23

varying levels of importance to the input variables when predicting the target variable.

The feature importance analysis revealed that Cd and Cr substantially impacted WQI, achieving higher relative importance compared to other variables. (AdB = 0.16, K-NN = 0.38, GB = 0.14, RF = 0.22, SVR = 0.40, and BR = 0.11) and (AdB = 0.10, K-NN = 0.38, GB = 0.31, RF = 0.13, SVR = 0.31, and BR = 0.27) respectively in all prediction models. In the K-NN and SVR models, the Hg variable demonstrated higher relative importance, with values of 1.89 and 2.06, respectively. However, for AdB and RF models, the Hg variable showed moderate importance, with values of 0.11 each. It is worth mentioning that there was substantial variation in the relative importance of the COD variable.

The analyses indicate that AdB, RF, and BR models displayed another comparable trend of superior relative importance for COD in predicting WQI. On the other hand, the K-NN model showed no significant importance for COD toward WQI prediction. A similar pattern can be observed for the NH_3–N variable. In the GB and RF models, NH_3–N demonstrated moderate relative importance, with values of 0.02 and 0.03, respectively. However, in the case of SVR and BR models, NH_3–N showed higher relative importance, with values of 1.32 and 1.03, respectively, indicating its significance in predicting the WQI. A notable finding is that, in most instances, the BOD variable showed a reduced relative importance compared to the other variables. The identification of the least significant variable by the models was that NO_3–N attained zero importance in GB, AdB, and RF models, SVR and BR highlighted TDS as the non-significant whereas K-NN and GB highlighted many other non-significant variables in the prediction of WQI (Fig. 9a–f & Table 13).

4. Discussion

There has been a growing interest in using machine-learning models to assess water quality, especially, in the recent years (Mondal et al., 2024). These models have the potential to detect changes quickly and effectively in water quality conditions (Huan and Liu, 2024). Our study aimed to enhance the field by assessing various machine learning models for the prediction of water quality. When comparing the predictive performance of the best models identified in our research, including Adaptive Boosting, K-Nearest Neighbors, Gradient Boosting, Random Forests, Support Vector Regression, and Bayesian Regression, the results indicate that the Gradient Boosting (GB) model, using sixteen input variables, outperforms other models in predicting the water quality index. While the remaining models produce similar results, they slightly lag the GB model in terms of predictive accuracy. Our findings are aligned with a study by (Khoi et al., 2022), found that boosting-based algorithms, especially Extreme Gradient Boosting (XGBoost), were highly accurate in predicting water quality for the La

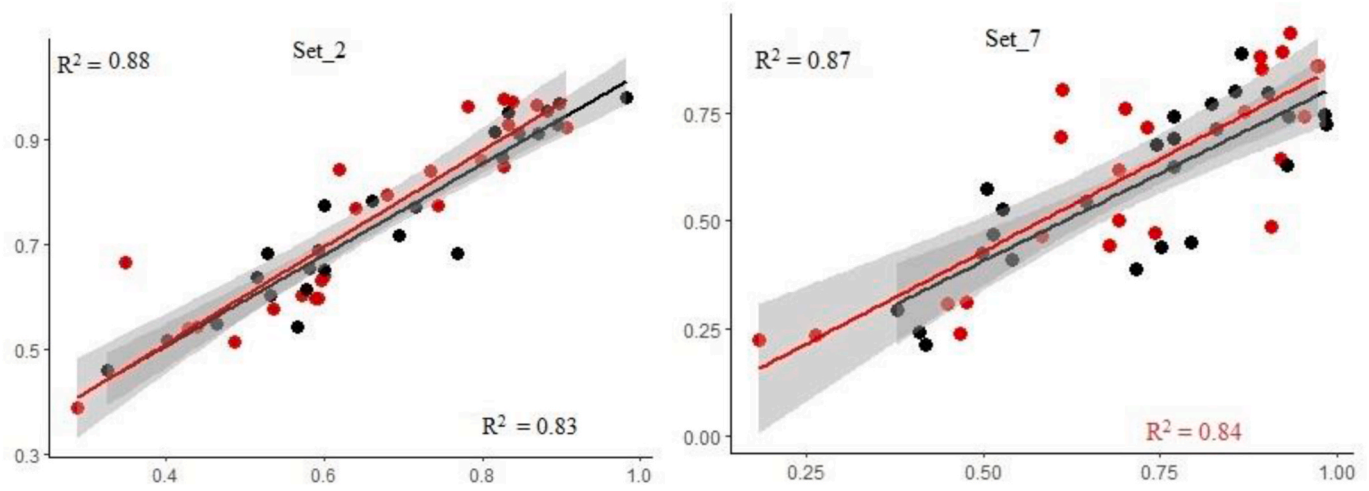

Fig. 7. a–b K-NN Algorithm's Best Inputs (2nd and 7th) for WQI Prediction (Training: Black, Testing: Red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 10
Evaluation measures for the SVR algorithm in WQI prediction for the aik-stream (training and testing datasets).

Model	R^2	RMSE	MAE	RAE (%)	RRSE (%)
Training					
1	0.74	4.75	5.3	61.71	106.35
2	0.67	8.35	3.39	90.22	120.77
3	0.73	6.34	3.91	71.13	57.34
4	0.65	3.56	5.71	72.23	117.21
5	0.45	6.67	3.99	71.11	77.31
6	**0.85**	**5.41**	**5.78**	**57.54**	**85.45**
7	0.43	9.72	3.76	74.23	84.34
8	0.63	3.35	3.312	50.21	94.67
9	0.25	3.23	3.51	47.25	81.57
10	0.71	4.34	4.64	66.78	59.98
11	0.63	9.23	3.32	90.23	76.32
12	0.55	4.56	5.21	69.42	73.32
13	0.43	4.38	4.56	44.55	86.12
14	0.71	9.12	3.37	74.22	53.32
15	0.55	4.14	4.78	41.22	85.24
16	0.65	5.43	5.23	44.99	109.12
17	0.54	4.57	4.34	71.34	68.21
18	0.73	3.47	5.56	48.33	101.33
19	0.54	7.34	4.77	34.23	94.44
Testing					
1	0.80	6.71	7.55	55.65	108.98
2	0.47	8.11	8.64	38.69	98.18
3	0.56	11.61	8.43	70.09	88.44
4	0.29	8.98	7.34	32.02	58.34
5	0.21	8.08	9.32	65.11	79.43
6	**0.82**	**6.67**	**7.23**	**71.54**	**79.23**
7	0.51	8.21	7.65	51.54	104.34
8	0.48	11.34	6.54	59.43	72.56
9	0.3	10.12	9.76	77.45	72.65
10	0.43	11.32	7.34	55.32	58.43
11	0.21	7.45	5.67	37.32	74.34
12	0.79	11.23	6.87	36.31	72.67
13	0.52	11.12	7.87	37.76	69.76
14	0.26	6.56	9.76	63.65	94.34
15	0.6	9.34	6.65	78.87	78.23
16	0.78	10.84	5.43	67.9	53.45
17	0.45	8.65	8.87	49.03	89.56
18	0.72	8.54	7.34	40.04	103.45
19	0.27	11.54	5.87	79.19	64.54

Buong River in Vietnam with high R^2 value of 0.989 and RMSE of 0.107. In a related study researchers have developed prediction models, for Water Quality Index using both Random Forest (RF) and Gradient Boosting (GB) algorithms at Rawal Lake, Pakistan (Ahmed et al., 2019). They discovered that the XGBoost model performed better than the RF model showing lower error rates (MAE = 1.9642, MSE = 7.2011, and RMSE = 2.6835) as compared to RF model (MAE = 2.3053, MSE = 9.5669 and RMSE = 3.0930). Many researchers have used the CCME-WQI as a crucial variable in their machine-learning models for predicting water quality trends (Yilma et al., 2018; Yu et al., 2020). Other notable research, such as the work of (Asadollah et al., 2021), highlighted the superiority of the Extra Tree Regression (ETR) model in their study that ETR outperformed other models with a high R-squared value of 0.97, demonstrating its effectiveness in predicting water quality while considering only ten variables.

Researchers in several countries have employed a variety of machine-learning models to predict water quality index, demonstrating high accuracy with R^2 values exceeding 0.90. For instance studies conducted by (Li et al., 2019) in Iraq focused on assessing water quality in the Euphrates River and in India (Nathan et al., 2017), evaluated groundwater quality using CCME-WQI in Lawspet, Puducherry and achieved an R^2 value above 90%. Likewise (Gazzaz et al., 2012) in Malaysia and researchers from Iran, including (Kamyab-Talesh et al., 2019) suggests that ML models are effective, at predicting and understanding water quality assessments.

Our findings support (Sakaa et al., 2022) study, which preferred Random Forest over Support Vector Regression (SVR) for predicting the Water Quality Index of Algeria's Wadi Saf-Saf river basin, achieving an R^2 of 0.82 and RMSE of 5.17 with thirteen parameters. However, Support Vector Regression (SVR) performed better than Artificial Neural Networks (ANN) in the study conducted by (Hazarika et al., 2020) in Tawang Chu River, Arunachal Pradesh, India. In a comparative study conducted by (Singha et al., 2021) on predicting groundwater quality in Arang, Chhattisgarh, India, also ranked XGBoost as the top-performing model, with R^2 = 0.962 in training and R^2 = 0.927 in testing phase followed by Artificial Neural Networks (ANN) and Random Forest (RF). Similarly, (Wong et al., 2022) conducted an extensive comparison of five regression models, including Multilayer Perception (MLP), Random Forest (RF), Decision Tree Regression (DTR), AdaBoost, and Support Vector Regression (SVR). According to their study the Random Forest algorithm outperformed other models, with a R squared value of 0.974 also aligns with our findings.

Besides ensemble methods like Random Forest and XGBoost, alternative machine learning algorithms, such as Bayesian Regularization by (Sakizadeh, 2015) and Artificial Neural Networks (ANN) by (Gazzaz et al., 2012), have also shown strong correlations between predicted and observed water quality values in their research. Comparative analyses among other ML models, different from the ones we mentioned, generally favored the Decision Tree model over algorithms like Naive

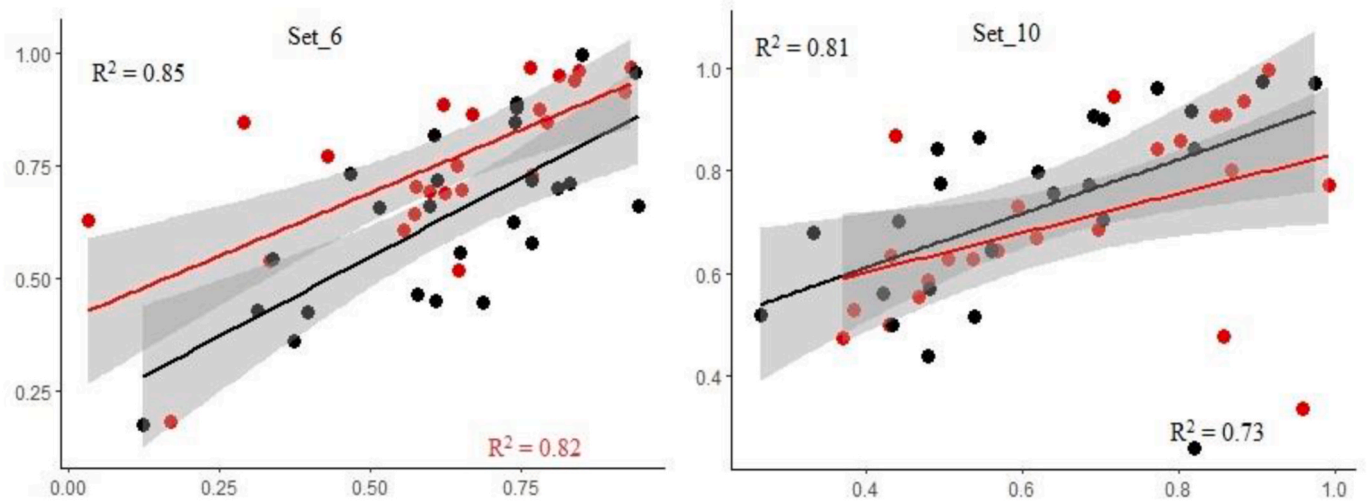

Fig. 8. a–b SVR Algorithm's Best Inputs (6th and 10th) for WQI Prediction (Training: Black, Testing: Red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 11
Findings from a regression analysis aimed at discovering the most suitable input combinations for modeling Water Quality Index (WQI) in the contaminated Aik-Stream.

No of parameters	Name of parameters	MSE	R^2	Adj R^2	AIC	BIC	Malkow PC
1	TDS	15.34	0.83	0.73	331.64	551.02	54.32
2	TDS, TSS	13.67	0.88	0.77	334.45	555.11	24.98
3	TDS, TSS, Pb	8.870	0.82	0.80	337.01	557.09	21.94
4	TDS, TSS, Pb, pH	11.56	0.82	0.81	337.67	559.34	12.09
5	TDS, TSS, Pb, pH, T	12.43	0.85	0.83	339.45	560.04	9.099
6	TDS, TSS, Pb, pH, T, Hg	14.83	0.85	0.83	340.69	562.87	11.32
7	TDS, TSS, Pb, pH, T, Hg, NO_3N	7.450	0.82	0.80	342.81	563.82	15.98
8	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD	7.321	0.87	0.88	344.91	564.82	23.12
9	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl	7.221	0.84	0.83	345.21	567.45	22.87
10	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr	6.898	0.84	0.83	346.41	568.32	21.54
11	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd	6.854	0.88	0.87	347.23	569.22	32.54
12	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD	7.342	0.88	0.86	347.88	572.67	11.54
13	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC	6.342	0.89	0.88	348.11	574.44	12.45
14	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N	6.811	0.89	0.87	349.03	574.81	11.32
15	**TDS, TSS, Pb,** pH, **T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N, Zn**	**4.034**	**0.93**	**0.92**	**351.54**	**579.42**	**10.54**
16	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N, Zn, As	6.231	0.90	0.87	352.22	576.43	18.87
17	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N, Zn, As, Ni	7.432	0.89	0.87	353.87	580.42	17.97
18	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N, Zn, As, Ni, Cu	5.213	0.88	0.84	353.81	581.33	15.74
19	TDS, TSS, Pb, pH, T, Hg, NO_3N, BOD, Cl, Cr, Cd, COD, TOC, NH_3N, Zn, Ni, As, Cu, OG	5.321	0.90	0.83	354.31	583.99	11.53

Table 12
Detailed examinations of multicollinearity statistics for the parameters used in the calculation of Water Quality Index (WQI).

Variable	VIF	1/VIF
COD	7.34	0.01239
TOC	9.47	0.042611
OG	9.15	0.06599
NH_3N	10.32	0.069834
As	11.43	0.087453
Ni	8.41	0.011895
Zn	7.98	0.0125269
Cd	6.47	0.0154488
Cr	5.95	0.0168013
Cl	4.97	0. 01308
BOD	4.36	0.029166
TDS	4.32	0.031383
TSS	4.31	0.031921
Pb	4.2	0.037852
pH	3.22	0.010983
T	3.2	0.012555
Hg	2.26	0.043095
NO_3N	1.93	0.018168
Cu	1.79	0.059737

Bayes (NB), k-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), and Logistic Regression (LogR) (M. Ahmed et al., 2021). Feature importance analysis by (Wong et al., 2022) showed that the proposed modified RF model, which included relatively important novel variables, is more proficient in water quality modeling. The predictive performance of the SVM model was observed to be reduced when irrelevant parameters were included in the input dataset (Leong et al., 2021). Therefore, it is noteworthy to select carefully and include only the relevant parameters to maximize the benefits of ML models in prediction tasks. Our research highlights that Gradient Boost and Random Forest models are the machine learning models with the strongest performance in achieving accurate predictive results in water quality modelling surpassing other prediction models found in existing literature. These models have proven their usefulness in various geographic contexts, providing valuable insights for water quality management and monitoring (Asadollah et al., 2021; Hibjur Rahaman, Roshani, Masroor and Sajjad, 2023). Researchers are continuously investigating novel algorithms and techniques to improve the accuracy and resilience of machine learning models (Ahmed et al., 2024; Khoi et al., 2022). Furthermore, they are actively developing approaches to transfer this new technique across various regions and water bodies, enabling the

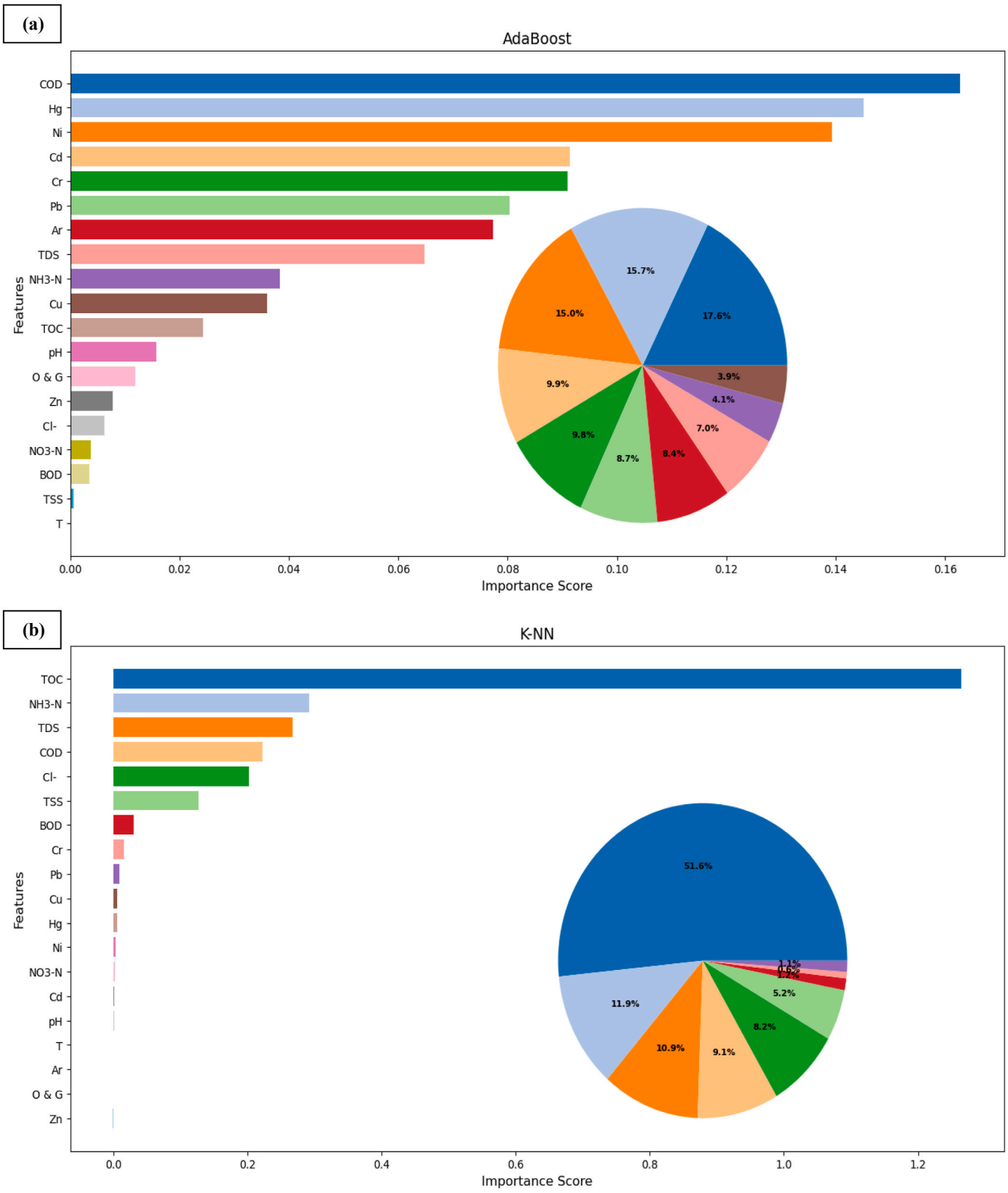

Fig. 9. a–f Feature importance scores of the predictive models (a)AdaBoost, (b) K-NN, (c) Gradient Boost, (d) Random Forest, (e) SVR, and (f) Bayesian regression.

utilization of these advanced techniques for better outcomes (Uddin et al., 2023).

5. Conclusion

This study evaluated the predictive performance of six distinct artificial intelligence (AI) models: AdaBoost, K-NN (K-Nearest Neighbors),

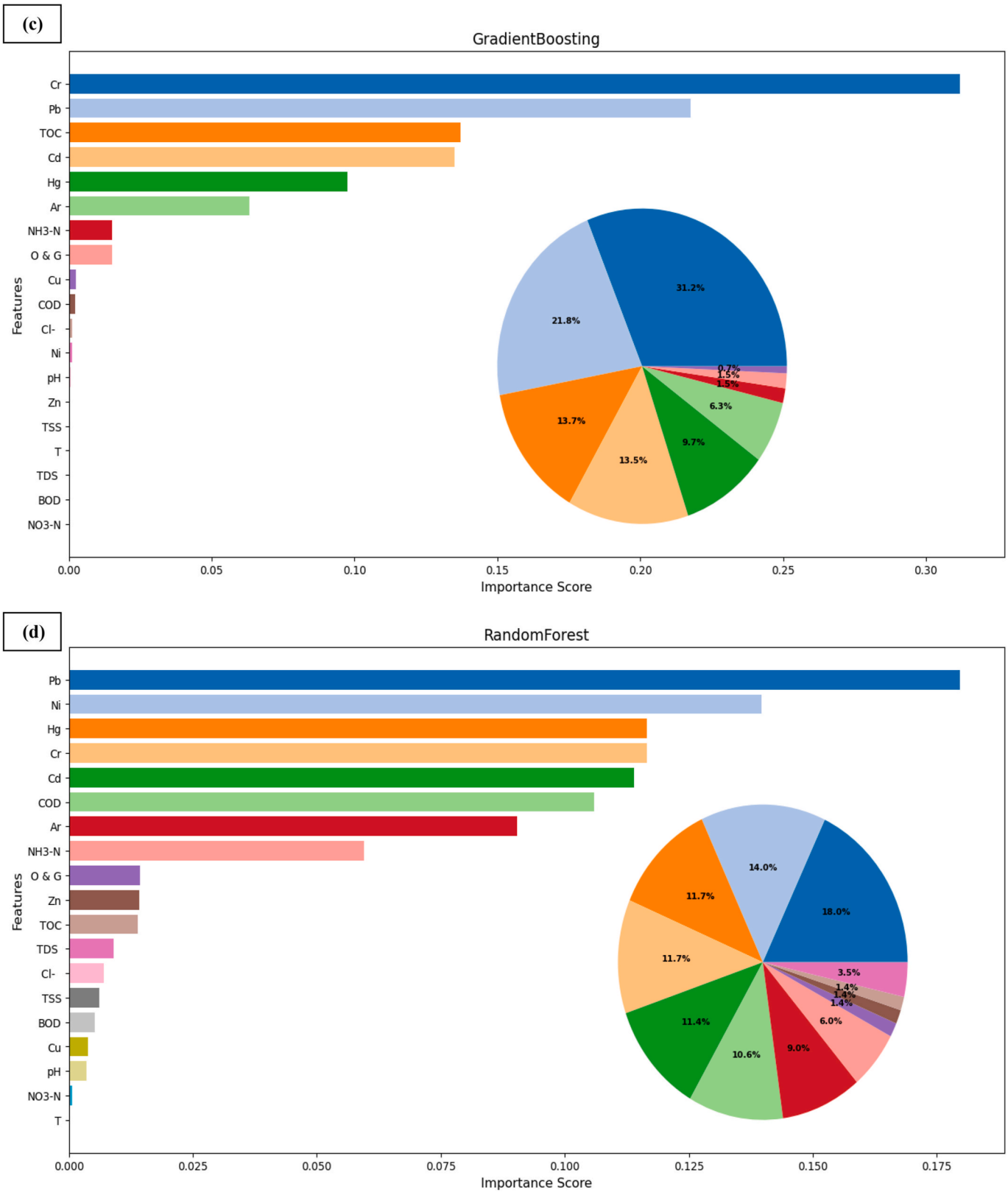

Fig. 9. (*continued*).

GB (Gradient Boosting), RF (Random Forests), SVR (Support Vector Regression), and BR (Bayesian Regression). The aimed was to forecast the water quality index of the Aik-Stream, utilizing data from 150 sampling locations measuring surface water quality. The study explored nineteen different combinations of input parameters, ranging from 1 to 19. The results indicated a notable decline in water quality, particularly in the midstream section of the Aik-Stream. The GB model demonstrated the highest accuracy achieving an R^2 of 0.88 (training) and 0.85 (testing) by utilizing only seven input variables. Although other models, like RF, BR, AdB KNN and SVR were less precise, they still offer

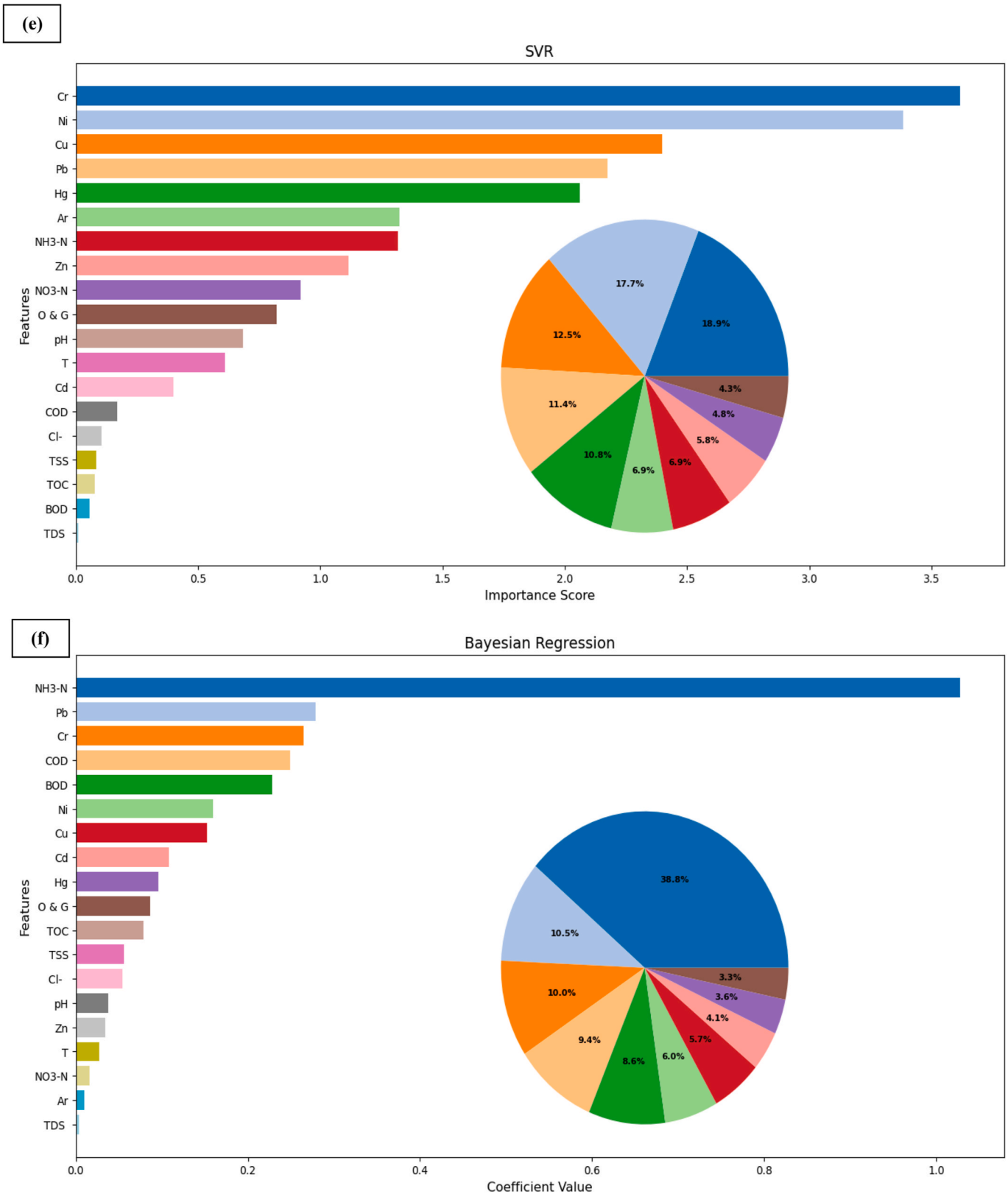

Fig. 9. (*continued*).

reasonably accurate predictions for the water quality index. All selected input variables significantly contributed to water quality index prediction, except for Nitrate Nitrogen (NO3–N), Total Dissolved Substances (TDS), and temperature (T) in this study.

Our findings suggest that the models examined in this research, particularly the GB model with the 13th input combination, have the potential to enable water managers and policymakers to efficiently calculate the water quality index for rivers and streams. This can be achieved with reduced computational time, lower costs, and less real-time monitoring at multiple polluted sites. Such an approach can

Table 13
Importance score of features in the predictive models.

Variables	Gradient Boost	Random Forest	AdaBoost	Support Vector	K-NN	Bayesian
pH	0.00	0.001	0.01	0.68	0.00	0.04
T C^0	0.00	0.00	0.00	0.61	0.00	0.03
COD mg/L	0.00	0.06	0.12	0.17	0.00	0.25
BOD mg/L	0.00	0.00	0.01	0.06	0.00	0.23
TDS mg/L	0.00	0.02	0.05	0.01	0.00	0.00
TSS mg/L	0.00	0.01	0.01	0.08	0.00	0.05
NH_3 -N mg/L	0.02	0.01	0.04	1.32	0.00	1.03
O& G mg/L	0.02	0.01	0.03	0.82	0.00	0.09
TOC mg/L	0.14	0.01	0.01	0.03	0.00	0.08
NO_3–N mg/L	0.00	0.01	0.00	1.32	0.01	0.06
Cl-mg/L	0.00	0.01	0.10	0.17	0.02	0.02
Cu mg/L	0.00	0.02	0.02	2.40	0.05	0.04
Zn mg/L	0.00	0.01	0.01	1.31	0.13	0.14
Cr mg/L	0.33	0.13	0.01	3.62	0.18	0.27
Pb mg/L	0.22	0.15	0.07	2.18	0.34	0.28
Cd mg/L	0.00	0.22	0.16	0.46	0.38	0.11
Ni mg/L	0.00	0.12	0.15	3.59	0.47	0.16
As mg/L	0.06	0.07	0.08	1.32	0.55	0.01
Hg mg/L	0.10	0.11	0.11	2.04	1.80	0.10

enhance water resource management strategies and lead to more effective assessments of water quality, ultimately contributing to sustainable river water management. However, it is important to acknowledge several limitations in this study, primarily related to the selection of physicochemical variables and the possibility of limited sampling sites. Future research should expand to include a larger number of sampling sites and incorporate a wider range of physicochemical variables to predict the river water quality index using an entropy weighted index approach. It is recommended to validate the proposed model in different rivers with varying hydro-climatic conditions.

CRediT authorship contribution statement

Ujala Ejaz: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Shujaul Mulk Khan:** Validation, Supervision, Project administration, checking the drafts of MS. **Sadia Jehangir:** Writing – review & editing, Investigation. **Noreen Khalid:** Writing – review & editing, Project administration. **Abdullah Abdullah:** Project administration, Investigation. **Majid Iqbal:** Formal analysis. **Zeeshan Ahmed:** Investigation & validation of the statistical analyses. **Aisha Nazir:** Funding acquisition. **Jens-Christian Svenning:** Writing – review & editing, Validation, Project administration, Funding acquisition.

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.

Data availability

Data will be made available on request.

Acknowledgments

Ujala Ejaz (first author) extends her gratitude to Higher Education Commission Pakistan, for supporting her studies in Denmark through the International Research Support Initiative Program (IRSIP). Furthermore, JCS considers this work a contribution to Center for Ecological Dynamics in a Novel Biosphere (ECONOVO), funded by Danish National Research Foundation (grant DNRF173) and his VILLUM Investigator project "Biodiversity Dynamics in a Changing World", funded by VILLUM FONDEN (grant 16549).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclepro.2024.141877.

References

Ahmed, A.A., Sayed, S., Abdoulhalik, A., Moutari, S., Oyedele, L., 2024. Applications of machine learning to water resources management: a review of present status and future opportunities. J. Clean. Prod. 441, 140715. https://doi.org/10.1016/j.jclepro.2024.140715.

Ahmed, M., Mumtaz, R., Zaidi, Hassan, M, S., 2021. Analysis of water quality indices and machine learning techniques for rating water pollution: a case study of Rawal Dam, Pakistan. Water Supply 21 (6), 3225–3250.

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A.A., Irfan, R., García-Nieto, J., 2019. Efficient water quality prediction using supervised machine learning. Water 11 (11), 2210.

Akhtar, F., Li, J., Pei, Y., Xu, Y., Rajput, A., Wang, Q., 2020. Optimal features subset selection for large for gestational age classification using gridsearch based recursive feature elimination with cross-validation scheme. Paper presented at the Frontier Computing: Theory, Technologies and Applications 8. FC 2019.

Ali, S., Zhang, S., Yue, T., 2020. Environmental and economic assessment of rainwater harvesting systems under five climatic conditions of Pakistan. J. Clean. Prod. 259, 120829.

Asadollah, S.B.H.S., Sharafati, A., Motta, D., Yaseen, Z.M., 2021. River water quality index prediction and uncertainty analysis: a comparative study of machine learning models. J. Environ. Chem. Eng. 9 (1), 104599.

Association, A.P.H., 1926. Standard Methods for the Examination of Water and Wastewater, vol. 6. American Public Health Association.

Bagherzadeh, F., Mehrani, M.-J., Basirifard, M., Roostaei, J., 2021. Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. J. Water Proc. Eng. 41, 102033.

Berglund, E.Z., Pesantez, J.E., Rasekh, A., Shafiee, M.E., Sela, L., Haxton, T., 2020. Review of modeling methodologies for managing water distribution security. J. Water Resour. Plann. Manag. 146 (8), 03120001.

Bhagat, S.K., Pilario, K.E., Babalola, O.E., Tiyasha, T., Yaqub, M., Onu, C.E., Yaseen, Z. M., 2023. Comprehensive review on machine learning methodologies for modeling dye removal processes in wastewater. J. Clean. Prod. 385, 135522. https://doi.org/10.1016/j.jclepro.2022.135522.

Bourel, M., Segura, A., 2018. Multiclass classification methods in ecology. Ecol. Indicat. 85, 1012–1021.

Brack, W., Dulio, V., Ågerstrand, M., Allan, I., Altenburger, R., Brinkmann, M., Escher, B. I., 2017. Towards the review of the European Union Water Framework Directive: recommendations for more efficient assessment and management of chemical contamination in European surface water resources. Sci. Total Environ. 576, 720–737.

CCME, 2001. Canadian Water Quality Guidelines for the Protection of Aquatic Life: CCME Water Quality Index 1.0, Technical Report: Canadian Council of Ministers of the Environment Winnipeg.

Chen, M., Ma, L.Q., 1998. Comparison of Four USEPA Digestion Methods for Trace Metal Analysis Using Certified and Florida Soils: Wiley Online Library.

Chen, R.-C., Dewi, C., Huang, S.-W., Caraka, R.E., 2020. Selecting critical features for data classification based on machine learning methods. Journal of Big Data 7 (1), 52.

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 Computer Science 7, e623.

Daily, D., 2006. No Step Taken to Check Tanneries Pollution, p. 12. March.

Danades, A., Pratama, D., Anggraini, D., Anggriani, D., 2016. Comparison of accuracy level K-nearest neighbor algorithm and support vector machine algorithm in classification water quality status. In: Paper Presented at the 2016 6th International Conference on System Engineering and Technology (ICSET).

Ebrahimi-Khusfi, Z., Nafarzadegan, A.R., Dargahian, F., 2021. Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques. Ecol. Indicat. 125, 107499.

Fida, M., Li, P., Wang, Y., Alam, S.K., Nsabimana, A., 2022. Water contamination and human health risks in Pakistan: a review. Exposure and Health 1–21.

Garai, J., 2014. Environmental aspects and health risks of leather tanning industry: a study in the Hazaribag area. Chinese Journal of Population Resources and Environment 12 (3), 278–282.

Gazzaz, N.M., Yusoff, M.K., Aris, A.Z., Juahir, H., Ramli, M.F., 2012. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar. Pollut. Bull. 64 (11), 2409–2420.

Geissen, V., Mol, H., Klumpp, E., Umlauf, G., Nadal, M., van der Ploeg, M., Ritsema, C.J., 2015. Emerging pollutants in the environment: a challenge for water resource management. International soil and water conservation research 3 (1), 57–65.
Gültekin, B., Sakar, B.E., 2018. Variable importance analysis in default prediction using machine learning techniques. In: Paper Presented at the 7th International Conference on Data Science, Technology and Applications, DATA 2018.
Hameed, M., Sharqi, S.S., Yaseen, Z.M., Afan, H.A., Hussain, A., Elshafie, A., 2017. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput. Appl. 28, 893–905.
Hazarika, B.B., Gupta, D., Ashu, Berlin, M., 2020. A comparative analysis of artificial neural network and support vector regression for river suspended sediment load prediction. In: Paper Presented at the First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019.
Hibjur Rahaman, M., Roshani, Masroor, M., Sajjad, H., 2023. Integrating remote sensing derived indices and machine learning algorithms for precise extraction of small surface water bodies in the lower Thoubal river watershed, India. J. Clean. Prod. 422, 138563. https://doi.org/10.1016/j.jclepro.2023.138563.
Hillel, T., Bierlaire, M., Elshafie, M.Z., Jin, Y., 2021. A systematic review of machine learning classification methodologies for modelling passenger mode choice. Journal of choice modelling 38, 100221.
Holmgren, N.M.A., Norrström, N., Aps, R., Kuikka, S., 2014. A concept of Bayesian regulation in fisheries management. PLoS One 9 (11), e111614.
Huan, S., Liu, X., 2024. Assessing the nonlinear relationship between consumer goods and water pollution in different seasons with machine learning models: a case study in the Yangtze River Economic Belt. J. Clean. Prod. 444, 141254. https://doi.org/10.1016/j.jclepro.2024.141254.
Hutton, M., Shafahi, M., 2019. Water Pollution Caused by Leather Industry: a Review. Paper Presented at the Energy Sustainability.
Islam, A.R.M.T., Talukdar, S., Mahato, S., Kundu, S., Eibek, K.U., Pham, Q.B., Linh, N.T. T., 2021. Flood susceptibility modelling using advanced ensemble machine learning models. Geosci. Front. 12 (3), 101075.
Issakhov, A., Alimbek, A., Zhandaulet, Y., 2021. The assessment of water pollution by chemical reaction products from the activities of industrial facilities: numerical study. J. Clean. Prod. 282, 125239. https://doi.org/10.1016/j.jclepro.2020.125239.
Jabeen, A., Huang, X., Aamir, M., 2015. The challenges of water pollution, threat to public health, flaws of water laws and policies in Pakistan. J. Water Resour. Protect. 7 (17), 1516.
Jadoon, W.A., Malik, R.N., 2019. Geochemical approach for heavy metals in suburban agricultural soils of Sialkot, Pakistan. SN Appl. Sci. 1 (2), 1–11.
Kamyab-Talesh, F., Mousavi, S.-F., Khaledian, M., Yousefi-Falakdehi, O., Norouzi-Masir, M., 2019. Prediction of water quality index by support vector machine: a case study in the Sefidrud Basin, Northern Iran. Water Resour. 46, 112–116.
Khalid, N., Rizvi, Z.F., Yousaf, N., Khan, S.M., Noman, A., Aqeel, M., Rafique, A., 2021. Rising metals concentration in the environment: a response to effluents of leather industries in Sialkot. Bull. Environ. Contam. Toxicol. 106, 493–500.
Khan, M.S.I., Islam, N., Uddin, J., Islam, S., Nasir, M.K., 2022. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences 34 (8), 4773–4781.
Khoi, D.N., Quan, N.T., Linh, D.Q., Nhi, P.T.T., Thuy, N.T.D., 2022. Using machine learning models for predicting the water quality index in the La Buong River, Vietnam. Water 14 (10), 1552.
Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artif. Intell. 97 (1–2), 273–324.
Kuhn, M., Johnson, K., 2018. Applied Predictive Modeling, 2nd 2018 Edition. Springer, March.
Landrigan, P., Fuller, R., Acosta, N., Adeyi, O., Arnold, R., Basu, N., 2017. Baldé, AB; Bertollini, R.; Bose: O'Reilly, S.
Leong, W.C., Bahadori, A., Zhang, J., Ahmad, Z., 2021. Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). Int. J. River Basin Manag. 19 (2), 149–156. https://doi.org/10.1080/15715124.2019.1628030.
Li, J., Abdulmohsin, H.A., Hasan, S.S., Kaiming, L., Al-Khateeb, B., Ghareb, M.I., Mohammed, M.N., 2019. Hybrid soft computing approach for determining water quality indicator: Euphrates River. Neural Comput. Appl. 31, 827–837.
Li, R.A., McDonald, J.A., Sathasivan, A., Khan, S.J., 2021. A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems. Water Res. 190, 116712.
Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., Wei, Y., 2013. A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Math. Comput. Model. 58 (3–4), 458–465.
Lokhande, R.S., Singare, P.U., Pimple, D.S., 2011. Toxicity study of heavy metals pollutants in waste water effluent samples collected from Taloja industrial estate of Mumbai, India. Resour. Environ. 1 (1), 13–19.
Mahmood, A., Syed, J.H., Malik, R.N., Zheng, Q., Cheng, Z., Li, J., Zhang, G., 2014. Polychlorinated biphenyls (PCBs) in air, soil, and cereal crops along the two tributaries of River Chenab, Pakistan: concentrations, distribution, and screening level risk assessment. Sci. Total Environ. 481, 596–604.
Malik, R.N., Jadoon, W.A., Husain, S.Z., 2010. Metal contamination of surface soils of industrial city Sialkot, Pakistan: a multivariate and GIS approach. Environ. Geochem. Health 32, 179–191.
Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using R for Digital Soil Mapping, vol. 35. Springer.
Maqbool, A., Ali, S., Rizwan, M., Ishaque, W., Rasool, N., Rehman, M.Z.U., Wu, L., 2018. Management of tannery wastewater for improving growth attributes and reducing chromium uptake in spinach through citric acid application. Environ. Sci. Pollut. Control Ser. 25, 10848–10856.
Mehdizadeh, S., Fathian, F., Safari, M.J.S., Khosravi, A., 2020. Developing novel hybrid models for estimation of daily soil temperature at various depths. Soil Tillage Res. 197, 104513.
Mienye, I.D., Sun, Y., 2022. A survey of ensemble learning: concepts, algorithms, applications, and prospects. IEEE Access 10, 99129–99149.
Modaresi, F., Araghinejad, S., 2014. A comparative assessment of support vector machines, probabilistic neural networks, and K-nearest neighbor algorithms for water quality classification. Water Resour. Manag. 28, 4095–4111.
Mokarram, M., Saber, A., Sheykhi, V., 2020. Effects of heavy metal contamination on river water quality due to release of industrial effluents. J. Clean. Prod. 277, 123380. https://doi.org/10.1016/j.jclepro.2020.123380.
Mokhtar, A., Jalali, M., He, H., Al-Ansari, N., Elbeltagi, A., Alsafadi, K., Rodrigo-Comino, J., 2021. Estimation of SPEI meteorological drought using machine learning algorithms. IEEE Access 9, 65503–65523.
Mondal, I., Hossain, S.K.A., Roy, S.K., Karmakar, J., Jose, F., De, T.K., Nguyen, N.-M., 2024. Assessing intra and interannual variability of water quality in the Sundarban mangrove dominated estuarine ecosystem using remote sensing and hybrid machine learning models. J. Clean. Prod. 442, 140889. https://doi.org/10.1016/j.jclepro.2024.140889.
Naeem, N., Khalid, N., Sarfraz, W., Ejaz, U., Yousaf, A., Rizvi, Z.F., Toxicology, 2021. Assessment of lead and cadmium pollution in soil and wild plants at different functional areas of Sialkot 107 (2), 336–342.
Nathan, N.S., Saravanane, R., Sundararajan, T., 2017. Application of ANN and MLR models on groundwater quality using CWQI at Lawspet, Puducherry in India. J. Geosci. Environ. Protect. 5 (3), 99.
Nguyen, Q.H., Ly, H.-B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Pham, B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Math. Probl Eng. 2021, 4832864. https://doi.org/10.1155/2021/4832864.
Pakistan, W., 2007. From water and health related issues in Pakistan. Fresh water and toxic programme 1–20.
Parween, S., Siddique, N.A., Diganta, M.T.M., Olbert, A.I., Uddin, M.G., 2022. Assessment of urban river water quality using modified NSF water quality index model at Siliguri city, West Bengal, India. Environmental and Sustainability Indicators 16, 100202.
Polikar, R., 2012. Ensemble learning. Ensemble machine learning: Methods and applications 1–34.
Qadir, A., Malik, R.N., 2009. Assessment of an index of biological integrity (IBI) to quantify the quality of two tributaries of river Chenab, Sialkot, Pakistan. Hydrobiologia 621, 127–153.
Qadir, A., Malik, R.N., Husain, S.Z., 2008. Spatio-temporal variations in water quality of Nullah Aik-tributary of the river Chenab, Pakistan. Environ. Monit. Assess. 140, 43–59.
Rabelo, L.M., Guimarães, A.T.B., de Souza, J.M., da Silva, W.A.M., de Oliveira Mendes, B., de Oliveira Ferreira, R., Malafaia, G., 2018. Correction to: histological liver chances in Swiss mice caused by tannery effluent. Environ. Sci. Pollut. Res. Int. 25 (16), 16267–16268.
Rigatti, S.J., 2017. Random forest. J. Insur. Med. 47 (1), 31–39.
Rodier, J., Bazin, C., Broutin, J., Chambon, P., Champsaur, H., Rodi, L., 2009. Water analysis, 9th edit. Dunod, Paris, France 1579.
Sakaa, B., Elbeltagi, A., Boudibi, S., Chaffaï, H., Islam, A.R.M.T., Kulimushi, L.C., Wong, Y.J., 2022. Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. Environ. Sci. Pollut. Control Ser. 29 (32), 48491–48508.
Sakizadeh, M., 2015. Assessment the performance of classification methods in water quality studies, A case study in Karaj River. Environ. Monit. Assess. 187, 1–12.
Sharma, A., Goyal, M.K., 2017. A comparison of three soft computing techniques, Bayesian regression, support vector regression, and wavelet regression, for monthly rainfall forecast. J. Intell. Syst. 26 (4), 641–655.
Siham, A., Sara, S., Abdellah, A., 2021. Feature selection based on machine learning for credit scoring: an evaluation of filter and embedded methods. In: Paper Presented at the 2021 International Conference on Innovations in Intelligent Systems and Applications (INISTA).
Singh, S., Jangir, S.K., Kumar, M., Verma, M., Kumar, S., Walia, T.S., Kamal, S., 2022. Feature importance score-based functional link artificial neural networks for breast cancer classification. BioMed Res. Int. 2022.
Singha, S., Pasupuleti, S., Singha, S.S., Singh, R., Kumar, S., 2021. Prediction of groundwater quality using efficient machine learning technique. Chemosphere 276, 130265. https://doi.org/10.1016/j.chemosphere.2021.130265.
Sun, A.Y., Scanlon, B.R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett. 14 (7), 73001.
Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M., 2020. Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data 7, 1–47.
Tariq, S.R., Shaheen, N., Khalique, A., Shah, M.H., 2010. Distribution, correlation, and source apportionment of selected metals in tannery effluents, related soils, and groundwater—a case study from Multan, Pakistan. Environ. Monit. Assess. 166, 303–312.
Teo, S.H., Ng, C.H., Islam, A., Abdulkareem-Alsultan, G., Joseph, C.G., Janaun, J., Awual, M.R., 2022. Sustainable toxic dyes removal with advanced materials for clean water production: a comprehensive review. J. Clean. Prod. 332, 130039. https://doi.org/10.1016/j.jclepro.2021.130039.

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.

Uddin, M.G., Nash, S., Rahman, A., Olbert, A.I., 2023. Performance analysis of the water quality index model for predicting water state using machine learning techniques. Process Saf. Environ. Protect. 169, 808–828.

Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., Zhang, H., 2021. Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environ. Res. 202, 111660.

Wang, Z., Luo, P., Zha, X., Xu, C., Kang, S., Zhou, M., Wang, Y., 2022. Overview assessment of risk evaluation and treatment technologies for heavy metal pollution of water and soil. J. Clean. Prod. 379, 134043. https://doi.org/10.1016/j.jclepro.2022.134043.

Water, U., 2017. Wastewater—The Untapped Resource; the United Nations World Water Development Report 2017. UNESCO.

Weiland, S., Hickmann, T., Lederer, M., Marquardt, J., Schwindenhammer, S., 2021. The 2030 Agenda for Sustainable Development: Transformative Change through the Sustainable Development Goals? Politics Gov. 9 (1), 90–95. https://doi.org/10.17645/pag.v9i1.4191.

Whitehead, P., Bussi, G., Hossain, M.A., Dolk, M., Das, P., Comber, S., Hossain, M.S., 2018. Restoring water quality in the polluted Turag-Tongi-Balu river system. In: Dhaka: Modelling Nutrient and Total Coliform Intervention Strategies. Science of the Total Environment, vol. 631, pp. 223–232.

Wong, W.Y., Al-Ani, A.K.I., Hasikin, K., Khairuddin, A.S.M., Razak, S.A., Hizaddin, H.F., Azizan, M.M., 2022. Water quality index using modified random forest technique: assessing novel input features. CMES-Computer Modeling in. Eng. Sci. 32 (3), 1011–1038.

Xu, Z., Cheng, L., Liu, P., Hou, Q., Cheng, S., Qin, S., Xia, J., 2022. Investigating the spatial variability of water security risk and its driving mechanisms in China using machine learning. J. Clean. Prod. 362, 132303. https://doi.org/10.1016/j.jclepro.2022.132303.

Yilma, M., Kiflie, Z., Windsperger, A., Gessese, N., 2018. Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia. Modeling Earth Systems and Environment 4, 175–187.

Yu, C., Yin, X., Li, H., Yang, Z., 2020. A hybrid water-quality-index and grey water footprint assessment approach for comprehensively evaluating water resources utilization considering multiple pollutants. J. Clean. Prod. 248, 119225. https://doi.org/10.1016/j.jclepro.2019.119225.

Zamani, M.G., Nikoo, M.R., Niknazar, F., Al-Rawas, G., Al-Wardy, M., Gandomi, A.H., 2023. A multi-model data fusion methodology for reservoir water quality based on machine learning algorithms and bayesian maximum entropy. J. Clean. Prod. 416, 137885. https://doi.org/10.1016/j.jclepro.2023.137885.

Zhang, J., Zou, T., Lai, Y., 2021. Novel method for industrial sewage outfall detection: water pollution monitoring based on web crawler and remote sensing interpretation techniques. J. Clean. Prod. 312, 127640. https://doi.org/10.1016/j.jclepro.2021.127640.

Zhou, Z., Chen, K., Li, X., Zhang, S., Wu, Y., Zhou, Y., Fan, W., 2020. Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. Nature Electronics 3 (9), 571–578.