IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 September 2023, accepted 12 September 2023, date of publication 14 September 2023,
date of current version 20 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3315649

RESEARCH ARTICLE

Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality

MUSHTAQUE AHMED RAHU[1], ABDUL FATTAH CHANDIO[1], KHURSHEED AURANGZEB[2], (Senior Member, IEEE), SARANG KARIM[3], MUSAED ALHUSSEIN[2], AND MUHAMMAD SHAHID ANWAR[4]

[1]Department of Electronic Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah 67450, Pakistan
[2]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, P. O. Box 51178, Riyadh 11543, Saudi Arabia
[3]Department of Telecommunication Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah 67450, Pakistan
[4]Department of AI and Software, Gachon University, Seongnam-si 13120, South Korea

Corresponding authors: Mushtaque Ahmed Rahu (marahu@quest.edu.pk) and Muhammad Shahid Anwar (shahidanwar786@gachon.ac.kr)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSPD2023R947.

ABSTRACT The degradation of water quality has become a critical concern worldwide, necessitating innovative approaches for monitoring and predicting water quality. This paper proposes an integrated framework that combines the Internet of Things (IoT) and machine learning paradigms for comprehensive water quality analysis and prediction. The IoT-enabled framework comprises four modules: sensing, coordinator, data processing, and decision. The IoT framework is equipped with temperature, pH, turbidity, and Total Dissolved Solids (TDS) sensors to collect the data from Rohri Canal, SBA, Pakistan. The acquired data is preprocessed and then analyzed using machine learning models to predict the Water Quality Index (WQI) and Water Quality Class (WQC). With this aim, we designed a machine learning-enabled framework for water quality analysis and prediction. Preprocessing steps such as data cleaning, normalization using the Z-score technique, correlation, and splitting are performed before applying machine learning models. Regression models: LSTM (Long Short-Term Memory), SVR (Support Vector Regression), MLP (Multilayer Perceptron) and NARNet (Nonlinear Autoregressive Network) are employed to predict the WQI, and classification models: SVM (Support Vector Machine), XGBoost (eXtreme Gradient Boosting), Decision Trees, and Random Forest are employed to predict the WQC. Before that, the Dataset used for evaluating machine learning models is split into two subsets: Dataset 1 and Dataset 2. Dataset 1 comprises 600 values for each parameter, while Dataset 2 includes the complete set of 6000 values for each parameter. This division enables comparison and evaluation of the models' performance. The results indicate that the MLP regression model has strong predictive performance with the lowest Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) values, along with the highest R-squared (0.93), indicating accurate and precise predictions. In contrast, the SVR model demonstrates weaker performance, evidenced by higher errors and a lower R-squared (0.73). Among classification algorithms, the Random Forest achieves the highest metrics: accuracy (0.91), precision (0.93), recall (0.92), and F1-score (0.91). It is also conceived that the machine learning models perform better when applied to datasets with

The associate editor coordinating the review of this manuscript and approving it for publication was Mouloud Denai.

smaller numbers of values compared to datasets with larger numbers of values. Moreover, comparisons with existing studies reveal this study's improved regression performance, with consistently lower errors and higher R-squared values. For classification, the Random Forest model outperforms others, with exceptional accuracy, precision, recall, and F1-score metrics.

INDEX TERMS Data collection, environmental monitoring, Internet of Things (IoT), machine learning, water quality analysis, water quality class (WQC), water quality index (WQI).

I. INTRODUCTION

In recent years, there has been a growing fascination with harnessing the capabilities of the Internet of Things (IoT) and machine learning paradigms for addressing environmental challenges. One such critical area of concern is water quality analysis and prediction. Accessing clean and safe water is a very fundamental requirement for mankind's health, agriculture, and ecosystem sustainability. However, deteriorating water quality due to pollution, population growth, and climate change has become a pressing issue worldwide. Traditional methods of water quality monitoring and prediction are often limited by their cost, time-consuming nature, and inability to capture real-time data. To overcome these limitations, the integration of IoT and machine learning technologies has emerged as a powerful solution. IoT enables the deployment of sensor networks in water bodies, collecting a vast amount of data on various Water Quality Parameters (WQPs) such as pH level, salinity, temperature, nutrients, and pollutant concentrations. This real-time data acquisition allows for continuous monitoring of water quality, offering a comprehensive understanding of the dynamic nature of aquatic ecosystems.

Fig. 1 illustrates the IoT-enabled water quality applications [1], such as aquaponics [2], aquaculture [3], fish ponds [4], water treatment [5], agriculture [6], [7], greenhouses [8], and irrigation [9]. IoT-enabled water quality systems can be intergraded with modern technologies, methods, and paradigms, that can achieve transformations and sustainable developments [10], for example, Artificial Intelligence (AI) (e.g., machine learning and deep learning) [11], [12], [13], big data [14], multifunctional sensors [15], [16], and renewable energy sources [17], [18].

Machine learning models provide powerful tools to analyze the collected data, identify patterns, and make accurate predictions [19], [20], [21]. By leveraging machine learning techniques, it becomes possible to develop models that can detect anomalies [22] and classify machine learning states [23]. This information can aid decision-makers, water resource managers, and policymakers in taking timely and informed actions to ensure the preservation and restoration of different resources of water [24]. Water quality monitoring in water bodies faces various environmental challenges that affect the monitoring procedures. Some of these challenges include pollution, sedimentation, seasonal variations, algal blooms and so on [25] and [26].

Addressing these environmental challenges requires a comprehensive approach that combines scientific expertise, stakeholder collaboration, and adequate resources. By overcoming these challenges, water quality monitoring efforts can provide valuable insights for effective water management and conservation in water bodies.

In this paper, we present frameworks based on IoT and machine learning to perform analysis and predict the water quality, with a specific focus on measuring temperature, pH, turbidity, and Total Dissolved Solids (TDS). These parameters are critical indicators of water quality, reflecting the physical and chemical characteristics of water bodies. The IoT component of our framework enables the deployment of sensor networks in water bodies, equipped with sensors for measuring the WQPs. These sensors collect data at regular intervals and transmit it wirelessly to a base station. This real-time data acquisition enables continuous monitoring of water quality, overcoming the limitations of traditional sampling methods. The IoT framework also facilitates remote access to the collected data, allowing water resource managers and stakeholders to monitor WQPs conveniently.

The collected data is then processed and analyzed using machine learning models within the framework. Two common types of machine learning models, for example regression models and classification models, are applied to derive insights from the collected data. The regression models and the classification models are employed for predicting the Water Quality Index (WQI) and Water Quality Class (WQC), respectively. By leveraging machine learning models, the framework provides actionable information that aids decision-makers and policymakers in making decisions and implementing proactive measures for water quality management and environmental conservation. The integration of these two technologies in water quality analysis and prediction offers a powerful tool for achieving efficient and sustainable water resource management. By enabling real-time monitoring, accurate analysis, and proactive decision-making, this framework contributes to the preservation and protection of water resources and ensures a safer and healthier environment for all.

The novelty of this study can be illustrated as follows:

- An integrated framework merging IoT and machine learning paradigms for comprehensive water quality analysis and prediction.
- Use of IoT-enabled sensors for real-time data collection from Rohri Canal, Shaheed Benazirabad (SBA), Pakistan, contributing to data accuracy and extensive inputs.
- Comparative analysis across distinct datasets (Dataset 1 and Dataset 2) offers insights into models'

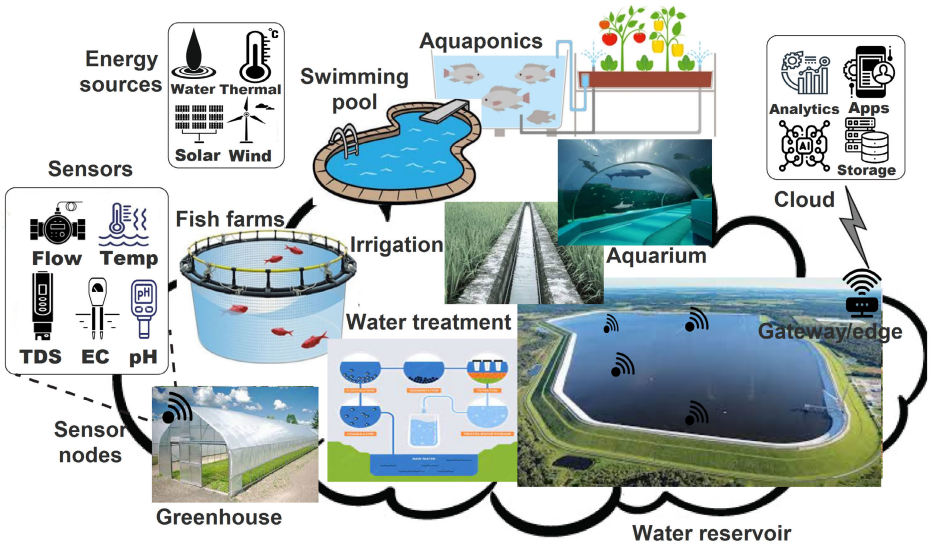


FIGURE 1. Internet of things-enabled water quality applications.

performance under varying data conditions, presenting empirical novelty.
- Examination of existing studies pertaining to water quality prediction, encompassing diverse methodologies, techniques, and models employed to predict WQI and WQC.

The remaining sections of the paper are structured as follows: Section II covers the existing literature and research related to the study. Section III discusses the material and methods, including the proposed IoT-enabled framework for data collection and the machine learning-enabled framework for data analysis. It also covers data preprocessing techniques and the machine learning models used, along with evaluation metrics. Section IV presents the performance evaluation of the machine learning models applied to water quality data analysis and prediction. In Section V, a brief discussion of the research findings is provided. Finally, Section VI concludes the paper by summarizing the findings.

II. RELATED WORK

Water quality analysis and prediction have been subjects of extensive research, with various studies exploring the integration of IoT and machine learning techniques for effective monitoring and management of water resources. The following is a review of relevant research and contributions in this field and summary of related in work is given in Table 1.

Kumar et al. [27] introduced an IoT-based infrastructure designed for monitoring and evaluating river water quality. The researchers conducted extensive experiments, collecting and analyzing water quality data from the Ganga River and Sangam River over different months and seasons. They collected continuous data samples for a duration of 15 months, by utilizing the Libelium smart water kit, equipped with sensors capable of measuring various parameters. After the data collection phase, the researchers utilized several models of machine learning to predict the water quality of both rivers. The study revealed that the water quality of both rivers was considered suitable for irrigation and fishing purposes. However, when assessing the average oxygen levels, it was concluded that the water was not suitable for use.

Tian et al. [28] performed a study on Sentinel-2 images to investigate and compare the performance of four machine learning models, namely eXtreme Gradient Boosting (XGBoost), Artificial Neural Network (ANN), Support Vector Regression (SVR), and Random Forest, in retrieving three WQPs for inland reservoirs. The study aimed to assess the effectiveness of these models in the context of water quality analysis. The results of the study demonstrated that XGBoost outperformed the remaining three models in accurately retrieving the WQPs. Building upon this finding, the researchers employed XGBoost to reconstruct the spatial-temporal patterns of the various parameters for the period spanning from 2018 to 2020. Moreover, they conducted a comprehensive analysis of the characteristics of interannual, seasonal, and spatial variation based on these reconstructed patterns. The study's outcomes provide a valuable and practical approach for monitoring and managing both optically and non-optically active WQPs at a regional scale. The use of machine learning models, particularly XGBoost, offers an efficient means of analyzing and interpreting water quality data from Sentinel-2 images.

Nasser et al. [29] conducted a study based on IoT technology, for which they deployed smart water meters to collect data at regular intervals, and the collected data was seamlessly transmitted to the cloud for storage and analysis. The authors developed a technology infrastructure utilizing microservices and containers to facilitate real-time streaming and enhance efficacy management. Machine learning approaches, specifically SVR and Random Forest were employed for time series forecasting applications. Through

a comparative study, the proposed model demonstrated superiority and served as a testing ground for other similar approaches.

Morón-López et al. [30] developed a remote monitoring system that centered around IoT technology and cloud-based data management. The authors conducted a comprehensive assessment of existing solutions, eventually selecting a personalized plug-and-play approach as their preferred method. To achieve continuous data transmission and retrieval, they deployed nodes that collected information and sent it to a web server. Subsequently, the gathered data was presented in real-time through a web interface, allowing users to visualize and analyze the water quality information conveniently. Additionally, the authors performed a Pearson correlation test to establish a relationship between the deployed nodes' data and satellite photographs. This correlation analysis aimed to augment the understanding of the accumulated data by the nodes and its relevance to satellite-based observations.

The big data analytics and data science fields are rapidly expanding, driven by increasing industrial demands. In a study by [31], a model was presented to identify the essential skills required for data science. The authors highlighted the significance of data skills and thoroughly examined the challenges associated with education in this domain. They reviewed several prominent projects that prioritize advanced data skills and also provided a use case demonstrating the application of frameworks in online learning management portals for developing big data skills.

Dong and Yan [32] proposed an effective solution for scheduling drainage, pumping, and water diversion in pumping stations using a data-driven model. They introduced a model predictive control system that utilizes supervised learning from IoT data and a short-term memory network model to simulate and predict water dynamics and flow quality. Through numerical analysis, the authors successfully demonstrated improved economic efficiency compared to existing benchmark solutions.

Wang et al. [33] emphasized the significance of wetlands as vital ecosystems for climate regulation and environmental protection. It highlights the adverse impact of human activities on wetland land cover, necessitating the use of remote sensing technology to monitor and classify land cover changes.The study investigates the efficacy of the Random Forest machine learning algorithm for classifying coastal wetland land cover using Worldview-2 and Landsat-8 imagery. A comparison with Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) methods reveals Random Forest's superior classification accuracy. Specifically, Random Forest achieves 91.86% accuracy for Worldview-2 images and 86.61% accuracy for Landsat-8 images. Even for limited-sample land cover types, Random Forest excels, and its performance is further enhanced by incorporating texture features. The research concludes that high-resolution remote sensing imagery favors small-scale land cover classification, and Random Forest emerges as the optimal choice for coastal wetland classification, surpassing SVM and K-NN algorithms.

Ghorbani et al. [34] have explored fluid-flow measurements in the petroleum industry, focusing on the relationship between orifice meter flow rate (Qv) and fluid-flow variables. The study employs five machine-learning algorithms to analyze data from an Iranian oil field pipeline. Correlations between Qv and variables like pressure, temperature, viscosity, square root of differential pressure, and oil specific gravity are evaluated. The five algorithms: Adaptive Neuro Fuzzy Inference System (ANFIS), Least Squares Support Vector Machine (LSSVM), Radial Basis Function (RBF), Multilayer Perceptron (MLP), and Gene Expression Programming (GEP) are tested on a dataset of 1037 records, with MLP providing the most accurate flow rate predictions, followed by GEP and RBF. ANFIS and LSSVM show lower accuracy, particularly at lower flow rates. The study suggests machine learning's potential to enhance orifice meter flow rate predictions, especially in challenging conditions, although further research on additional datasets is necessary for confirmation.

A. MOTIVATION AND SCOPE

IoT and machine learning [35] are both technologies that possess the potential to be transformative and impactful in various fields, including water quality monitoring, but they differ in their approach and the type of data they use. IoT [36] involves numerous sensors and devices to accumulate and transmit data over the internet. IoT can be used to monitor a wide range of WQPs, like temperature, pH, dissolved oxygen, and more.IoT sensors can be strategically deployed throughout water treatment plants, distribution systems, and other locations to provide real-time data on water quality.

In contrast, machine learning [37] is a subset of AI that empowers machines (e.g., computers, robots, vehicles) to acquire knowledge from data and utilize it to make predictions and decisions. Machine learning algorithms can analyze large and complex datasets to detect patterns and relationships that may be difficult for humans to identify, resulting in more accurate predictions of water quality. Machine learning can be used to develop predictive models for WQPs, allowing for early detection of issues. But machine learning models is preferred for smaller datasets. Whereas, to handle large datasets, deep learning models are usually adopted. Deep learning is the subset of machine learning models, specifically tailored to address the constraints and limitations posed by traditional machine learning models when grappling with extensive datasets, often referred to as ''big data'' [38]. Additionally, machine learning models that typically necessitate ongoing supervision and re-training, in contrast to deep learning networks often require minimal or no additional training after initial calibration.

Henceforth, both IoT and machine learning [39] can be used for water quality monitoring, IoT focuses on collecting and transmitting data from sensors and devices,

TABLE 1. Summary of related work.

Study	Focus	Approach and techniques	Dataset/method	Findings and contributions
Kumar et al. [27]	IoT-based river water quality monitoring	Data collection using Libelium smart water kit; machine learning models (e.g., SVM, ANN) applied to Ganga River and Sangam River data	Water quality data collected over 15 months; machine learning models used for prediction	Machine learning models predict water quality; water deemed unsuitable for certain uses based on oxygen levels
Tian et al. [28]	Machine learning models for water quality analysis	Four machine learning models (XGBoost, ANN, SVR, Random Forest) evaluated on Sentinel-2 images; WQP retrieval	Sentinel-2 images; Comparison of machine learning models; XGBoost outperforms others	XGBoost efficiently analyzes optically and non-optically active WQPs from Sentinel-2 images
Nasser et al. [29]	Real-time water data analysis with IoT	Deployment of smart water meters; Microservices and containers for data analysis; machine learning models (SVR, RF) for forecasting	IoT-based data collection; Real-time data streaming and forecasting using machine learning models	Proposed model excels in real-time streaming and forecasting with IoT data
Morón-López et al. [30]	IoT-based cloud data management	IoT nodes deployed for continuous data transmission; Real-time data visualization through a web interface; Correlation analysis with satellite photos	IoT nodes data; Correlation analysis with satellite observations	Real-time water quality monitoring and correlation with satellite data
Menasalvas et al. [31]	Essential skills for data science	Identification of key data science skills; Examination of challenges in data science education	Analysis of data science skill requirements; Discussion on addressing education challenges	Framework for developing data science skills and overcoming education challenges
Dong and Yang [32]	Data-driven model for water dynamics	IoT data from pumping stations; Short-term memory network model for predictive control	Pumping stations data; Data-driven model for drainage, pumping, and water diversion	Improved economic efficiency in water management using data-driven model
Wang et al. [33]	Random Forest algorithm for wetland classification	Utilization of Random Forest, SVM, K-NN on high-resolution imagery (Worldview-2, Landsat-8); Texture features incorporated	High-resolution imagery analysis; Comparison of Random Forest, SVM, K-NN algorithms	Random Forest excels in coastal wetland classification using high-resolution imagery
Ghorbani et al. [34]	Flow rate prediction through wellhead chokes	Comparison of five machine learning algorithms (ANFIS, LSSVM, RBF, MLP, GEP) on Iranian oil field data	Iranian oil field data; Comparative study of machine learning algorithms	MLP provides accurate flow rate predictions; Potential for enhanced predictions using machine learning algorithms
This study	An integrated framework that combines IoT and machine learning to analyze and predict water quality	Utilizes IoT framework for data collection; Four regression models are used to predict WQI and four classification models are used to predict the WQC	Data collection involves IoT sensors for temperature, pH, turbidity, and TDS from Rohri Canal, SBA, Pakistan; Dataset is bifurcated into two subsets: Dataset 1 (600 values) and Dataset 2 (6000 values) to enable a comprehensive evaluation of model performance	This work not only proposes a novel framework for water quality analysis and prediction but also provides practical insights for robust monitoring strategies

while machine learning focuses on analyzing data to make predictions and decisions. Both technologies can be used together to provide a comprehensive approach to water quality monitoring and management. The features, limitations and applications of IoT can be found in [40] and for machine learning in [41] and [42].

III. MATERIALS AND METHODS

A. STUDY AREA: ROHRI CANAL, SBA

Fig. 2 shows the study area; which is Rohri Canal, SBA, Pakistan. Fig. 2b illustrates the Google Map of the water quality monitoring site and Fig. 2c illustrates the water quality monitoring field stations.

Rohri Canal is located in Nawabshah City, SBA District of Sindh province in Pakistan, and serves as a significant waterway in the region. Drawing water from the Indus River near Sukkur, the Rohri Canal acts as a gravity canal, utilizing the natural flow of water due to the elevation difference. It forms an extensive network of distributors and minor canals, ensuring that water reaches various areas and agricultural fields within the SBA District. Rohri Canal is an important water resource for the region, providing irrigation water for agriculture and serving as a drinking water source for local communities. Water quality monitoring assumes a vital role in the evaluation and preservation of the quality of water in Rohri Canal, ensuring its suitability for various purposes. The monitoring process involves measuring various physical, chemical, and biological parameters. Samples are collected regularly from different locations along the canal and analyzed in laboratories using standard protocols. The monitoring frequency depends on factors like the significance of the water source, regulatory requirements, and available resources.

In response to the challenging manual process of collecting water samples from Rohri Canal, we recommend an IoT-based system that could monitor the water quality of the canal. The proposed system would utilize interconnected devices and sensors deployed along the canal to automate data collection. The sensors would continuously measure parameters like temperature, pH, turbidity, and TDS. The collected data would be transmitted wirelessly to a central unit for real-time analysis and storage. The proposed

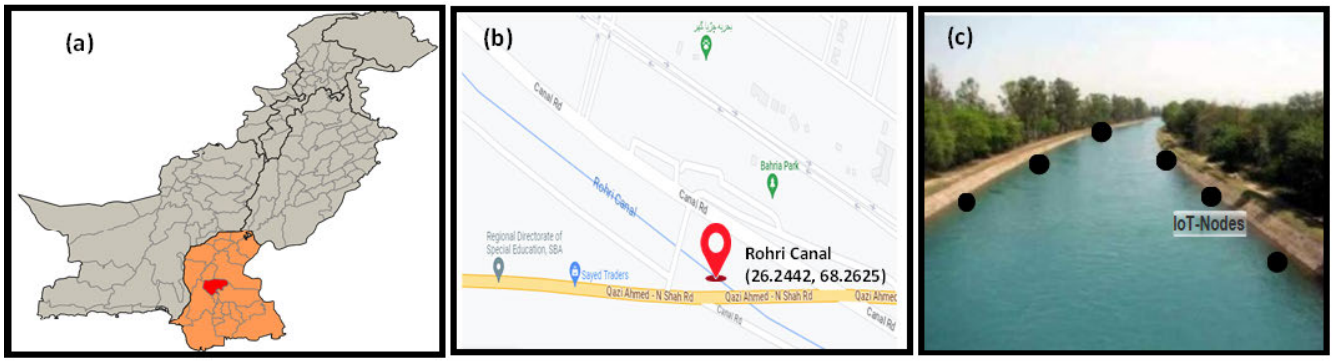


FIGURE 2. Geographic locations of the study area in (a) Shaheed Benazirabad, Sindh, Pakistan [Courtesy: Wikipedia] (b) site [Courtesy: Google Maps] (c) stations.

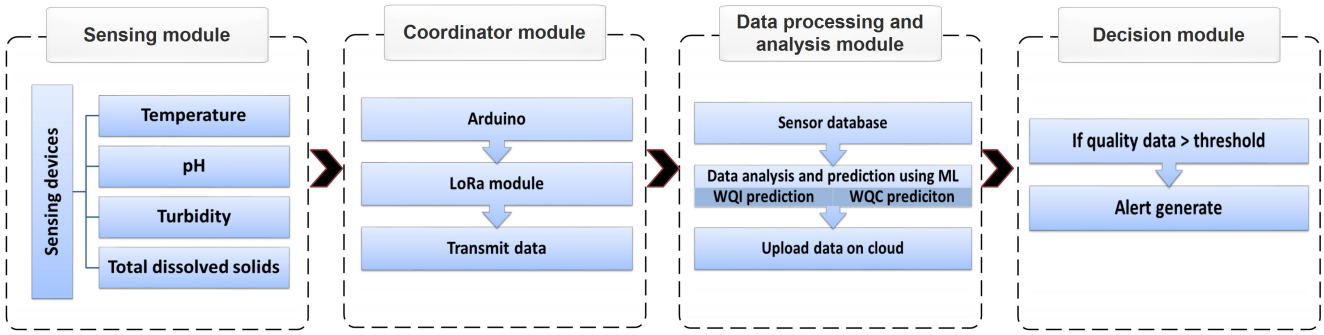


FIGURE 3. IoT-enabled framework for water quality.

TABLE 2. Hardware list.

Station	Hardware	Model
Base station	LoRa gateway	SX1278 Lora transceiver Module
	Laptop / Computer	Any
IoT node	Arduino controller	Arduino Uno R3
	LoRa	SX1278 Lora transceiver Module
	Temperature sensor	DS18B20 Temperature Sensor
	pH sensor	SEN0161 pH meter
	Turbidity sensor	SEN0189 Analog Turbidity Sensor
	TDS sensor	KS0429 TDS Meter V1.0

IoT-enabled system would provide timely and accurate information on water quality, enabling prompt action and effective monitoring of the canal's water conditions.

B. IoT-ENABLED FRAMEWORK FOR WATER QUALITY

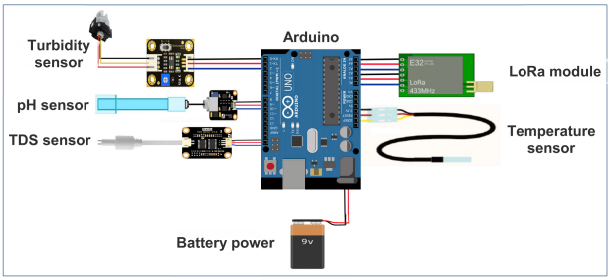
Fig. 3 depicts the proposed IoT-enabled framework for water quality. The connection diagram of our IoT-enabled system is depicted in Fig. 4, in which Fig. 4a and Fig. 4b are connection diagrams for an IoT node and a base station, respectively. The list of hardware, along with model number and quantity, is populated in Table 2. The proposed IoT-enabled framework has four modules, such as sensing module, coordinator module, data processing module, and decision module. The sensing module consists of temperature, pH, turbidity, and TDS sensors. The coordinator module consists of an Arduino controller and Long-Range (LoRa). The data processing module consists of a database, a machine learning block, and uploading the data to the cloud. The decision module generates an alert if the water quality is beyond the threshold. The tasks for each module are discussed as follows.

1) SENSING MODULE

The sensing module is responsible for collecting data on WQPs. In this case, it consists of temperature, pH, turbidity, and TDS sensors. These sensors are deployed in the water bodies to periodically measure the respective WQPs. The temperature sensor measures the water temperature, the pH sensor measures the acidity or alkalinity, the turbidity sensor measures the clarity or cloudiness, and the TDS sensor measures the concentration of dissolved solids. These sensors provide real-time data on the WQPs, they consist of an Arduino with a LoRa module and four sensors. The connection diagram of an IoT is shown in Fig. 4a. The real-time IoT node (or sensor node) at the deployment location is depicted in Fig. 5.

a: TEMPERATURE SENSOR

These sensors [43] are used to measure the temperature of water bodies in water quality monitoring systems. They provide critical information about the thermal characteristics of the water, which is important for understanding the



(a) IoT node

(b) Base station

FIGURE 4. Connection diagram of proposed IoT-enabled system.

behaviour of aquatic organisms, chemical reactions, and overall ecosystem dynamics. Temperature sensors can be based on different technologies, such as resistance temperature detectors, thermocouples, or thermistors. They are typically designed to be waterproof and resistant to corrosion, allowing them to function accurately in aquatic environments. Temperature data collected by these sensors help to assess the impacts of climate change, monitor water quality in industrial processes, and ensure the suitability of water for various applications.

b: pH SENSOR

This sensor measures the acidity or alkalinity of water by quantifying the concentration of hydrogen ions in the water [44]. They provide a pH value that ranges from 0 to 14, with values below 7 considered acidic, values above 7 considered alkaline, and a pH of 7 representing neutrality. The pH sensors employ various technologies, such as glass electrodes or solid-state sensors, to measure pH accurately. They are essential in water quality monitoring as pH affects biological processes, chemical reactions, and the overall balance of aquatic ecosystems. pH sensors are used in applications such as drinking water analysis, aquaculture, wastewater treatment, and industrial processes where maintaining a specific pH range is critical.

c: TURBIDITY SENSOR

This sensor measures the clarity or cloudiness of water caused by suspended particles [45]. They work by emitting light into the water and measuring the scattering and absorption of the light as it interacts with the particles. Turbidity is expressed in Nephelometric Turbidity Units (NTU). Turbidity sensors are important in water quality monitoring as they indicate the presence of sediments, suspended solids, or pollutants. They are broadly employed in environmental monitoring, treatment plants for drinking water, and wastewater management systems to assess water quality, detect changes in turbidity levels, and identify potential issues affecting aquatic ecosystems and human health.

d: TOTAL DISSOLVED SOLIDS (TDS) SENSOR

TDS sensors measure the total concentration of dissolved solids in water [46]. They detect and quantify the presence

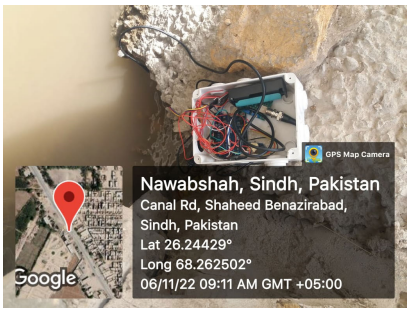


FIGURE 5. Real-time IoT node at deployment location.

of dissolved substances, such as minerals, salts, metals, and other organic and inorganic compounds. TDS is commonly represented in milligrams per liter (mg/L) or parts per million (ppm). TDS sensors operate based on different principles, including conductivity or optical sensors. They provide insights into the overall purity and mineral content of water. Monitoring TDS levels is important for various applications such as drinking water analysis, hydroponics, industrial processes, and boiler feedwater treatment. By measuring TDS, potential issues related to water quality and the accumulation of harmful substances can be identified, and appropriate measures can be taken to ensure the safety and suitability of the water.

e: SENSOR ADVANTAGES AND COST-EFFECTIVE OPTIMIZATION FOR WATER QUALITY ANALYSIS

The selected sensors, namely temperature, pH, turbidity, and TDS, offer distinct advantages in water quality analysis, collectively contributing to a comprehensive understanding of aquatic ecosystems [47], [48]. Temperature sensors provide real-time information about water temperature variations, which can be crucial for assessing the health of aquatic life and identifying potential thermal pollution. pH sensors enable precise monitoring of acidity or alkalinity levels, aiding in the detection of water bodies' potential vulnerability to pollutants and changes in natural processes. Turbidity sensors play a pivotal role in gauging water clarity, helping identify sediment levels, suspended particles, and potential contaminants that can influence overall water quality. TDS

sensors quantify the concentration of dissolved solids, such as salts and minerals, which is instrumental in assessing water's suitability for specific purposes, from drinking to industrial use.

One of the notable advantages of using these sensors is their potential for cost-effectiveness. The implementation of IoT technology enables real-time data collection and transmission, reducing the need for frequent on-site monitoring and manual data collection. This automation leads to optimized resource utilization, as personnel can focus on analysis and decision-making rather than spending extensive time on data collection.

Optimization techniques further enhance the cost-effectiveness of sensor deployments. Leveraging optimization algorithms, such as genetic algorithms or particle swarm optimization, can aid in determining the optimal locations for sensor placement [49], [50]. This strategic positioning ensures maximum coverage and accuracy while minimizing the number of sensors required. Moreover, these techniques can optimize sensor operation schedules, minimizing energy consumption and prolonging sensor lifespans.

By combining these advantages of sensor technology and optimization techniques, the study not only provides a comprehensive water quality analysis but also offers an economically viable solution. This approach reduces operational costs, enhances data accuracy, and supports sustainable resource management. The integration of sensors, IoT technology, and optimization methodologies demonstrates the potential to revolutionize water quality monitoring by making it both technically efficient and economically feasible.

2) COORDINATOR MODULE

The coordinator module assumes the responsibility of coordinating and managing various tasks or components within a system, for example, it connects the sensing module with the data processing module. It consists of an Arduino controller and LoRa technology. The Arduino controller acts as the central processing unit that collects the data from the sensors in the sensing module. LoRa technology is used for long-range wireless communication, allowing the Arduino controller to transmit the collected data to the next module. The integration of the Arduino controller with LoRa and sensors is depicted in Fig. 4a.

a: ARDUINO CONTROLLER

This is a versatile microcontroller board that acts as the central processing unit in various IoT applications including water quality monitoring [51]. The Arduino controller provides a platform for programming and interfacing with sensors, actuators, and other components. It offers a wide range of input and output pins, allowing for easy connectivity with external devices. Regarding water quality monitoring, the Arduino controller assumes a pivotal role in the coordination and management of data collection from sensors. It receives data from sensors and performs necessary computations or manipulations on the data if required. The Arduino controller acts as a bridge between the sensing module and the data processing module, facilitating the communication and transfer of data to further stages of the system.

b: LoRa

It is a low-power, long-range wireless communication technology that enables efficient and reliable data transmission over long distances [52], [53]. It is particularly well-suited for IoT applications, including water quality monitoring systems. LoRa technology operates in the license-free radio spectrum, providing excellent penetration through obstacles and long-range connectivity. LoRa technology is used in the coordinator module to disseminate the collected data from the Arduino controller to the data processing module. It establishes a wireless connection and eliminating the need for wired connections. LoRa's long-range capability allows the data to be transmitted over extended distances, making it suitable for remote or distributed monitoring systems. With LoRa, water quality data can be reliably transmitted from remote locations to the data processing module, enabling real-time monitoring and analysis.

3) DATA PROCESSING AND ANALYSIS MODULE

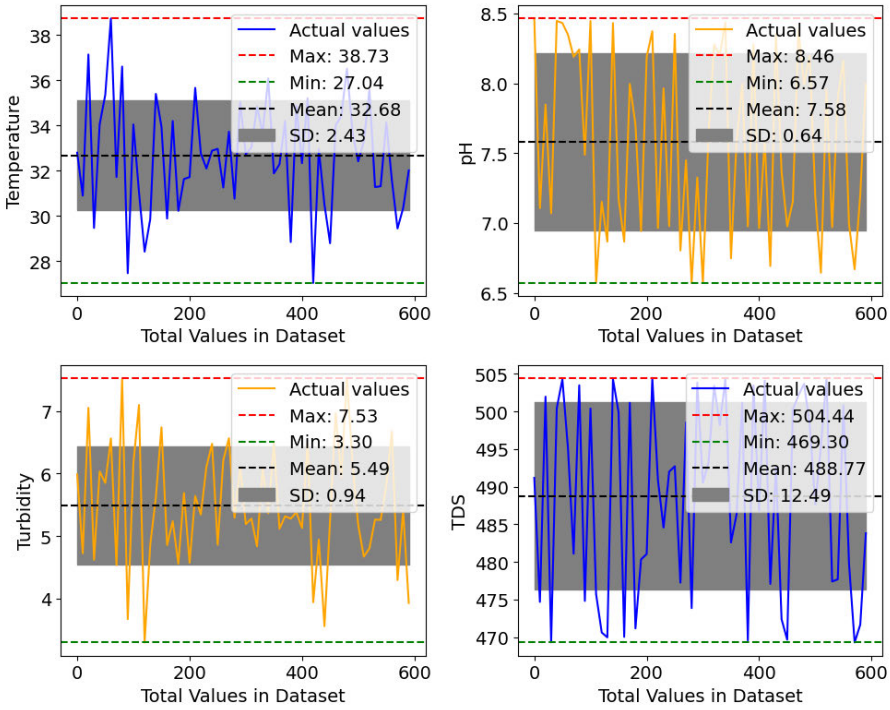
This module receives the data from the coordinator module and processes it for further analysis and storage. It consists of three components: a database, a machine learning block, and uploading the data to the cloud. The database keeps the collected water quality data for future reference and analysis. The machine learning block employs machine learning models to analyze the data and extract insights and patterns from it. This can help in predicting the anomalies and potential water quality issues. The processed data, along with relevant information, is then uploaded to the cloud for storage and accessibility.

4) DECISION MODULE

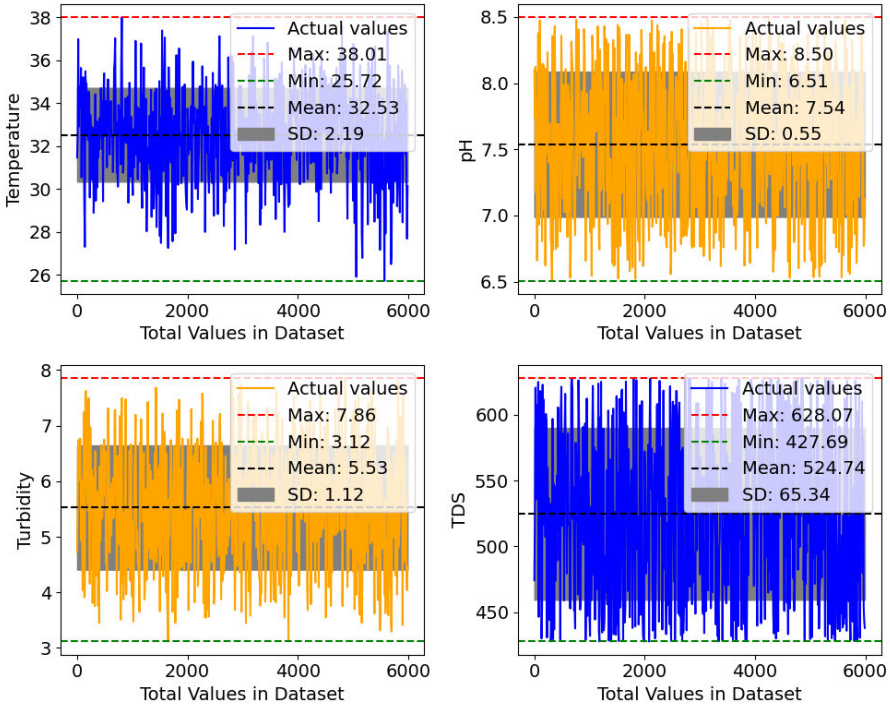
The decision module is responsible for generating alerts or notifications if the water quality exceeds certain predefined thresholds or if specific conditions are met. It evaluates the processed data and compares it against the threshold values set for each parameter. If the water quality data exceeds the thresholds, indicating a potential problem, the decision module generates an alert. The alert can be in the form of a notification sent to stakeholders or a visual indication on a monitoring dashboard. This allows for timely intervention and corrective actions to be taken to address the water quality issue.

C. DATASET GENERATION

This section discusses the process of generating the Dataset, which is collected from Rohri Canal, SBA using an IoT framework. The first step is to strategically deploy



(a) Dataset 1



(b) Dataset 2

FIGURE 6. **Statistical analysis of water quality parameters of datasets.**

sensors in water bodies. These sensors include temperature, pH, turbidity, and TDS sensors. Each sensor is carefully placed to capture specific WQPs of Rohri Canal, SBA. For example, temperature sensors may be placed at different depths to capture temperature variations within the water column, while pH and turbidity sensors are positioned to measure acidity/alkalinity and clarity/cloudiness, respectively.

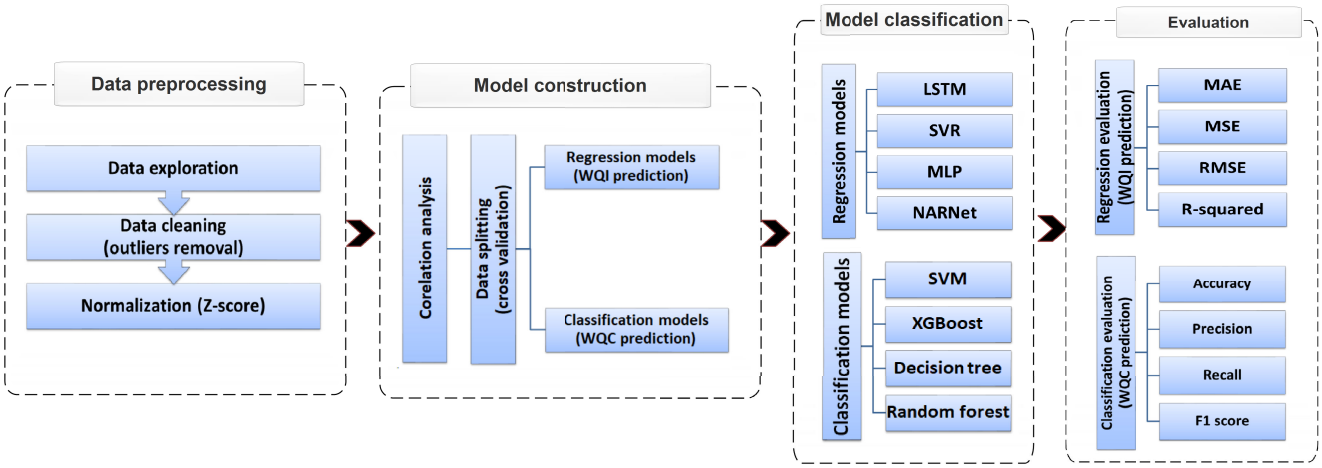


FIGURE 7. Machine learning-enabled framework for water quality.

Once the sensors were deployed, they performed continuous real-time measurements of the targeted WQPs. For instance, temperature sensors were responsible for tracking water temperature fluctuations, pH sensors quantified hydrogen ion concentrations, turbidity sensors gauged particle cloudiness, and TDS sensors evaluated dissolved solids content. These sensors transmitted data at regular intervals to a central base station using wireless communication. Subsequently, the base station stored the accumulated data and managed a dedicated database. A total of 6000 data points were collected for each parameter, creating a comprehensive dataset for subsequent analysis. This meticulous data collection approach, undertaken over the course of four months during the Autumn Season and limited to daytime hours for security, ensured the reliability and representativeness of the dataset for robust analysis and prediction.

To assess the efficacy of the machine learning models, the Dataset is partitioned into two subsets: Dataset 1 and Dataset 2. Dataset 1 contains 600 values for each parameter, while Dataset 2 contains the full 6000 values for each parameter. This division allows for comparative analysis and performance evaluation of the models.

The datasets are visualized in Fig. 6a and Fig. 6b, respectively, representing the data distribution. Where x-axis indicates the total values in the Dataset, while the y-axis indicates the corresponding water quality parameter values. The figures also depict the maximum level, minimum level, mean level, and standard deviation (SD) of the parameter values.

These statistical measures, including the maximum level, minimum level, mean level, and SD, help in understanding the characteristics and variability of the water quality parameters in the Dataset. They offer valuable insights into the data's range and distribution, which are essential for further analysis and interpretation.

D. MACHINE LEARNING-ENABLED FRAMEWORK FOR WATER QUALITY

The proposed machine learning-enabled framework for water quality analysis is depicted in Fig. 7. The input data is collected from Rohri Canal, SBA using the IoT framework. The collected data is preprocessed and cleaned. Later, the data is normalized using the Z-score technique, correlated and split, respectively. These are some preliminary steps to be performed before applying the machine learning models. After the preliminary steps, several machine learning models are utilized to analyze and predict the WQI and WQC indices. We calculate the WQI using four regression models, such as Long Short-Term Memory (LSTM), SVR, MLP, and Nonlinear AutoRegressive Network (NARNet) and calculate the WQC using four classification models, such as SVM, XGBoost, Decision Tree and Random Forest. The regression models are evaluated and analysed using Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) error metrics. Whereas, classification models are evaluated and analyzed using accuracy, precision, recall, and F1 score metrics. Based on these performance metrics, the results of regression models and classification models are demonstrated, tabulated, and compared accordingly.

1) DATA PREPROCESSING

The processing phase holds great significance in data analysis as it plays a pivotal role in enhancing the quality of the data. The acceptable limits for WQPs recommended by the World Health Organization (WHO) for safe drinking water and irrigation purposes are mentioned in [54] and [55] and are enlisted in Table 3. The data is collected through the IoT system and later is cleaned. After data cleaning, the WQI is calculated using the most significant parameters. Subsequently, water samples have been categorized or classified according to their corresponding WQI values.

TABLE 3. The acceptable limits for WQPs set by the WHO [54], [55].

Parameters	Acceptable limits		Unit
	Drinking water	Irrigation	
Temperature	< 25	< 25	°C
pH	6.5 - 8.5	6.0 - 7.5	n/a
Turbidity	1.0 - 5.0	n/a	NTU
TDS	600	< 2000	mg/L

To achieve improved accuracy, the Z-score method has been employed as a data normalization technique. The following section discusses different phases of data preprocessing.

a: DATA EXPLORATION

The data used for this research is gathered with the aid of IoT devices as discussed earlier. The Dataset has four significant WQPs, namely, temperature, pH, turbidity, and TDS. It contained about 6000 values (i.e., 6000 readings for each parameter), collected during the Autumn season. Later, it is divided into two subsets for machine learning model assessment, i.e., Dataset 1 (600 values/parameters) and Dataset 2 (complete set).

b: DATA CLEANING (OUTLIERS REMOVAL)

Data cleaning, specifically outliers removal, is an important step in the data preprocessing phase of a machine learning framework [56]. Outliers are data points that exhibit substantial deviation from the remaining data, indicating their significant divergence from the norm, and they can distort statistical analysis and model training. Outlier removal aims to identify and handle these anomalous data points for improving the quality and reliability of the datasets. In this study, we collected data using IoT devices. As a result, the datasets contain very few missing values and thus have a minimal number of outliers. To identify outliers, we employed the boxplot technique, as illustrated in Fig. 8. The outliers in Dataset 1 and Dataset 2 are illustrated in Fig. 8a and Fig. 8b, respectively. Dataset 1 contains fewer outliers compared to Dataset 2 because Dataset 2 is larger in size than Dataset 1.

c: WATER QUALITY INDEX (WQI)

WQI is a comprehensive metric that provides a singular measure to assess the overall quality of water. It is computed by considering a range of parameters that accurately reflect the true quality of the water [57]. To conventionally calculate the WQI, four WQPs are used, namely temperature, pH, turbidity and TDS in our datasets. Utilizing the assigned weights for each parameter, we computed the WQI for each sample using the formula shown in Equation 1. In this equation, q_{value} represents the value of a specific parameter within the range of 0-100, and w_{factor} denotes the weight coefficients associated with each parameter, as illustrated in Fig. 9. WQI is obtained by summing the products of the q value and the respective weight for each parameter. This sum is then divided by the total sum of the weights assigned to the parameters [58], [59].

$$WQI = \frac{(\sum q_{value} + w_{factor})}{\sum w_{factor}} \tag{1}$$

d: WATER QUALITY CLASS (WQC)

Water Quality Classes (WQCes) are used to assess and describe the condition or suitability of water for specific purposes, such as drinking, recreational activities, or ecosystem health [60]. The classes are typically based on the measurements of various parameters, and indicators of water quality, such as WQI range values. The specific definition of WQCes may vary as per context and standards or guidelines used by different organizations or regulatory bodies. These classes are often determined by setting thresholds or ranges for key WQPs and assigning a class label based on the measurements falling within those ranges. Based on the WQI range values, the WQCes can be defined as follows [24]:

- **Excellent:** Water that meets or exceeds all regulatory standards for drinking water quality.
- **Good:** Water that meets most of the regulatory standards but may have slight variations in certain parameters.
- **Fair:** Water that meets some of the regulatory standards but may require additional treatment or monitoring.
- **Marginal:** Water that fails to meet several regulatory standards and requires significant treatment or remediation.
- **Poor:** Water that is not suitable for drinking and irrigation due to severe contamination or the presence of harmful substances.

Hence, utilizing the WQI, water is classified into different categories as demonstrated in Fig. 10. The Y-axis describes the WQI value, and the X-axis corresponds to the WQC.

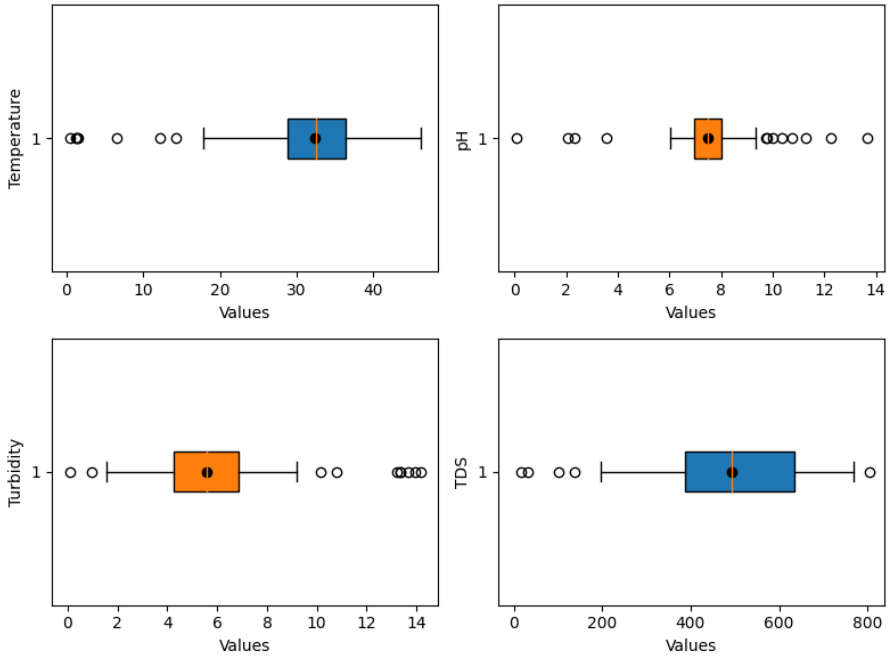
e: Z-SCORE NORMALIZATION

Data standardization of variables is a crucial step before training the machine learning model. This method is commonly employed in machine learning techniques to transform all data variables into a consistent scale for optimizing the training errors [61], [62]. In the current study, the data was normalized using the Z-score normalization method. Z-score normalization is a widely used method that normalizes parameters based on the mean (μ) and standard deviation (σ) values of the observed data. It is calculated using Equation 2, where x describes the value of a specific sample [63].

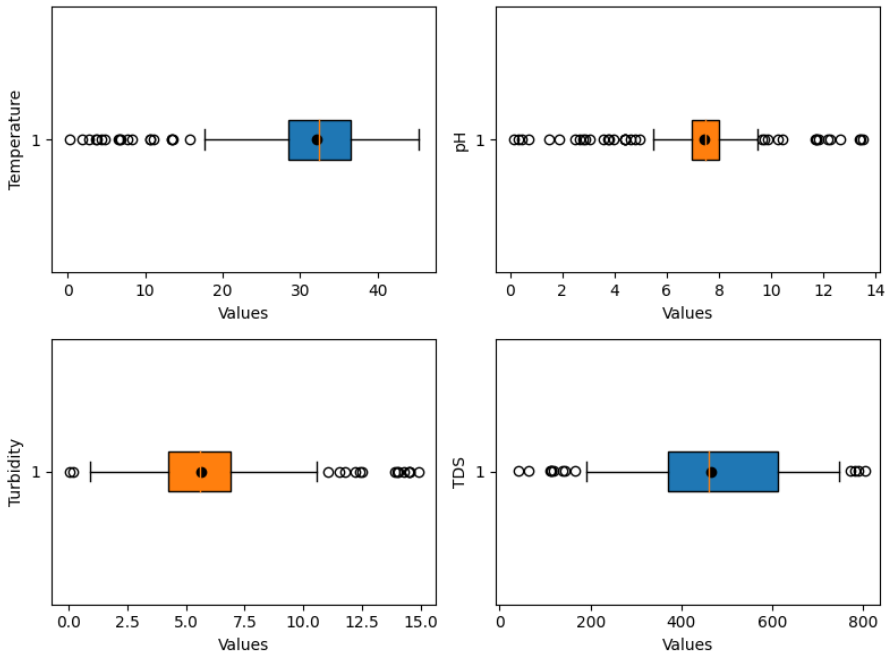
$$Z - score = \frac{(x + \mu)}{\sigma} \tag{2}$$

2) DATA ANALYSIS

Following the completion of data processing, multiple machine learning models are utilized for data analysis. The objective was to predict the target variables (WQI and WQC) using a minimal set of parameters. Before



(a) Dataset 1



(b) Dataset 2

FIGURE 8. **Outliers detection using boxplot analysis.**

predicting the target variables, certain preliminary steps were undertaken to set the data for input into the models. These steps included correlation analysis to identify relationships between variables and data splitting to partition the Dataset appropriately. These preparatory measures were crucial in ensuring the data was well-prepared for subsequent analysis with the machine learning models.

a: CORRELATION ANALYSIS

Correlation analysis, a statistical method, was employed to examine relationships between variables in the datasets. This analysis played a crucial role in understanding the degree and direction of associations between the variables [64]. In water quality analysis, correlation analysis can help identify and understand the relationships between different WQPs. The

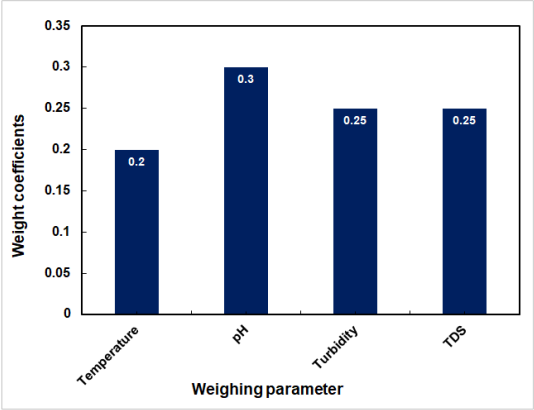


FIGURE 9. Weight coefficients used in the calculation of the WQI.

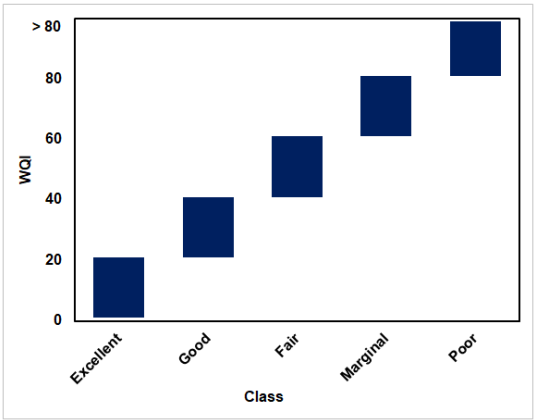


FIGURE 10. WQI range values [24].

correlation coefficient is a numerical value that ranges from −1 to 1 [65]. Here's what the correlation coefficient values represent:

- A correlation coefficient near to 1 suggests a strong positive correlation, indicating that as one parameter increases, the other parameter also tends to increase.
- A correlation coefficient near to −1 suggests a strong negative correlation, suggesting that as one parameter increases, the other parameter tends to decrease.
- A correlation coefficient near to 0 suggests a weak or no correlation, suggesting that there is no significant correlation between the parameters.

By analyzing the correlation coefficients, we can gain insights into how different WQPs influence each other. For example, a positive correlation between temperature and pH might indicate that as temperature increases, the pH tends to increase as well. This information can be useful for understanding water quality trends, identifying potential impacts on ecosystems, and even predicting the behaviour of certain parameters based on others. Fig. 11 represents the correlation analysis for the four parameters such as temperature, pH, turbidity, and TDS in terms of the heatmap diagram. Accordingly, Fig. 11a and Fig. 11b show the correlation analysis for Dataset 1 and Dataset 2, respectively. For correlation analysis, we employed the effective and widely used correlation method and is Pearson correlation method. It is a well-established statistical technique that measures the relationship between two variables [66].

b: DATA SPLITTING (CROSS-VALIDATION)

Before applying the machine learning model, the final task involves dividing the given data into separate sets to facilitate model training, testing, and performance evaluation [67]. Cross-validation is a statistical method utilized for evaluating the efficacy of a machine learning model. This method entails dividing the available data into multiple subsets or folds. The general procedure of cross-validation can be summarized as follows [68]:

- The Dataset is partitioned into k-folds having identical sizes.
- In each iteration of k-fold cross-validation, the model is trained using k-1 folds as training data, and the remaining fold is used for validation.
- Performance metrics, such as accuracy, precision, recall, or F1 score, are computed for each iteration based on the evaluation results from the validation set.
- The performance scores from all iterations are averaged to obtain an overall performance estimate of the model.
- The model parameters can be fine-tuned based on the average performance score, and the final model can be trained on the entire Dataset using the selected parameters.

Typically, popular choices for cross-validation include 5-fold and 10-fold [69]. Our study implemented 5-fold cross-validation, which involved dividing the data into 5 subsets or folds.

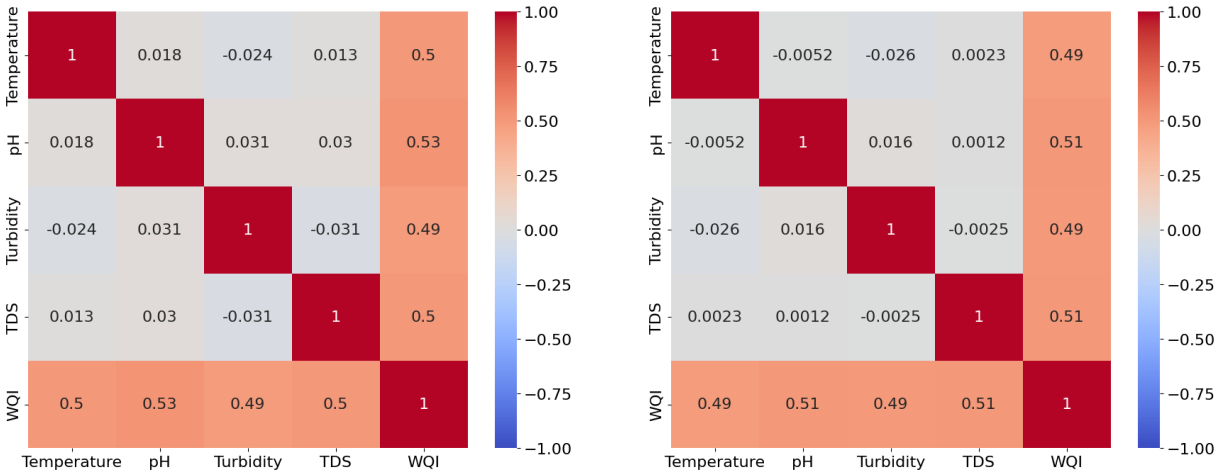
3) MACHINE LEARNING MODELS

In our study, we employed both regression and classification models of machine learning. Regression models were utilized to estimate the WQI, while classification models were employed to classify samples into predefined WQC. We employed a total of eight machine learning models out of which four are regression models and four classification models. In the subsequent section, we offer a comprehensive explanation of the models utilized in our analysis.

a: REGRESSION MODELS FOR WQI PREDICTION

For this purpose, LSTM, SVR, MLP and NARNet models are used for the prediction of WQI.

(i) LSTM Model: The LSTM model [70], categorized under the Recurrent Neural Network (RNN) architecture, is specifically engineered to capture long-term dependencies and manage sequential data. Their effectiveness in handling sequential data and capturing long-term dependencies has made them a popular choice for these tasks. One key distinction of LSTMs from traditional RNNs is their capability to mitigate the vanishing gradient problem that often arises



(a) Dataset 1 (b) Dataset 2

FIGURE 11. Pearson's correlation analysis using heatmap diagram.

during the training of deep NNs. LSTMs achieve this by incorporating memory cells and gates (input gate, output gate, and forget gate) that control the flow of information. The memory cells store and update information over time, allowing the model to retain relevant context and remember important information from the past. The gates, which regulate the flow of information, facilitate the selective retention or omission of information. This way, LSTMs can effectively capture long-term dependencies and avoid losing important information over extended sequences. Here are the control parameters of LSTM model:

- Number of hidden LSTM units
- Dropout rate
- Learning rate
- Batch size
- Number of epochs

(ii) SVR Model: The SVR model [71] is a specialized variant of SVM adapted for regression tasks. Its primary objective is to discover a hyperplane that optimally fits the training data while minimizing the deviations between predicted values and actual targets. SVR employs support vectors, which are data points in close proximity to the hyperplane, to establish a margin that encompasses the regression function. By utilizing a kernel function, the data is transformed into a higher-dimensional space, enabling SVR to handle both linear and nonlinear regression problems effectively. This method proves particularly valuable when dealing with datasets that exhibit intricate relationships. Following are the control parameters of SVR model:

- Kernel type (linear, polynomial, radial basis function, etc.)
- Kernel parameters (e.g., gamma for RBF kernel)
- Regularization parameter (C)
- Epsilon for epsilon-insensitive loss
- Degree for polynomial kernel

(iii) MLP Model: The MLP model, on the other hand, is a type of ANN, featuring multiple layers of interconnected nodes, commonly referred to as neurons. Within this structure, each neuron processes the weighted sum of its inputs, applying a non-linear activation function. This ability enables the MLP to grasp intricate patterns and relationships present in the data. MLP is a versatile algorithm and can be applied to various tasks, including regression [72]. It is known for its ability to approximate any continuous function given sufficient data and training time. MLP models can incorporate multiple hidden layers, and the performance of the model can be enhanced by adjusting the number of neurons in each layer. This flexibility allows for fine-tuning the architecture to achieve optimal results based on the specific problem and dataset at hand. The control parameters of MLP model are given below:

- Number of hidden layers
- Number of neurons in each hidden layer
- Activation functions for each layer
- Learning rate
- Batch size
- Number of epochs

(iv) NARNet Model: The NARNet model is a type of NN architecture specifically designed for time series forecasting tasks. NARNet combines the power of autoregressive models and neural networks to capture the nonlinear dependencies and dynamics present in sequential data. Unlike traditional autoregressive models, NARNet usually employs neural networks, which allows it to capture complex patterns and nonlinearity in the time series data. The model takes a fixed-length input window of past observations and uses a set of hidden layers to map these inputs to the predicted output. NARNet can be trained using gradient-based optimization algorithms, such as backpropagation, to minimize the prediction error. NARNet is applied in various domains, where

accurate and nonlinear forecasting is crucial. By leveraging the power of neural networks, NARNet provides a flexible and effective approach for time series forecasting tasks [73]. The control parameters of NARNet models are as follows:

- Number of lag observations
- Number of neurons in hidden layers
- Learning rate
- Batch size
- Number of epochs

b: CLASSIFICATION MODELS FOR WQC PREDICATION

For this purpose, SVM, XGBoost, Decision Trees, and Random Forest models are used for the prediction of WQC.

(i) SVM Model: The SVM model [74] is a powerful and versatile machine learning model that is widely used for both regression and classification purposes. SVMs are particularly effective when dealing with complex datasets and problems with a clear separation between classes. SVMs can handle both linear and non-linear separable data by utilizing different types of kernels, such as linear, polynomial, radial basis function, and sigmoid kernels, which allow for mapping the data into higher-dimensional feature spaces. SVMs have the advantage of being less prone to overfitting and providing robust generalization. They are also effective in handling high-dimensional data and can handle datasets with a large number of features. SVMs have found applications in various domains, such as bioinformatics, text classification, and image recognition. Their ability to handle complex decision boundaries and generalize well makes SVMs a valuable tool in the field of machine learning. Now, we enlist different control parameters of SVM model:

- Kernel type (linear, polynomial, radial basis function, etc.)
- Kernel parameters (e.g., gamma for RBF kernel)
- Regularization parameter (C)

(ii) XGBoost Model: XGBoost [75] is an advanced machine learning model that has gained significant popularity for its exceptional performance and versatility. It belongs to the gradient-boosting family and is known for its ability to handle complex datasets and deliver highly accurate predictions. XGBoost combines the power of ensemble learning with gradient boosting, creating a robust model by sequentially adding weak learners (usually Decision Trees) to correct the errors of the previous models. To prevent overfitting and enhance the generalization ability of the model, XGBoost employs regularization methods such as L1 and L2 regularizations. XGBoost optimizes the objective function through gradient-based optimization, efficiently computing the gradient and second derivatives of the loss function. Additionally, it offers features like handling missing values, feature importance estimation, and parallel processing to enhance efficiency. Following are control parameters of XGBoost model:

- Number of boosting rounds
- Learning rate (eta)
- Maximum depth of trees
- Minimum child weight
- Subsample ratio of the training instances
- Column subsample ratio for each tree
- Regularization term (lambda)
- Gamma parameter for regularization
- Scale_pos_weight for handling class imbalance

(iii) Decision Tree Model: The Decision Tree model is a versatile machine learning model that leverages a tree-like structure to make predictions by following a series of decision rules. It is known for its simplicity and effectiveness and finds applications in both regression and classification tasks. Decision tree models are extensively employed in tackling machine learning tasks owing to their inherent simplicity and comprehensibility. This popularity is justified by their ability to offer intuitive insights into the decision-making process underlying complex data patterns [76]. The Decision Tree model starts with a root node and recursively splits the Dataset on the basis of values of input features, creating branches and leaf nodes. Each internal node represents a decision based on a feature, while each leaf node represents a predicted class or value. The splitting process is determined by criteria such as Gini impurity or information gain, aiming to minimize the impurity or maximize the information gain at each step. Decision Trees are highly interpretable, as they allow us to trace the path of decisions leading to a prediction. Decision Trees are useful for feature selection, as they provide insights into the most important features for prediction. Nevertheless, there is a potential risk of overfitting when the Decision Tree becomes overly complex. To mitigate this concern, several techniques can be utilized, including pruning and the utilization of ensemble methods such as Random Forests, which can effectively address the issue of overfitting [77]. Here are control parameters of Decision Tree model:

- Splitting criterion (e.g., Gini impurity, entropy)
- Maximum depth of the tree
- Minimum number of samples required to split an internal node
- Minimum number of samples required to be at a leaf node

(iv) Random Forest Model: Random Forest [78] is a versatile and robust machine learning model that combines the power of multiple Decision Trees to make accurate predictions. It belongs to the ensemble learning methods and is known for its efficiency, scalability, and ability to handle complex datasets. In a Random Forest model, a group of Decision Trees is trained independently on distinct subsets of the training data, employing a technique called bootstrapping, which involves random sampling with replacement. Each tree in the ensemble casts a vote for the final prediction, and the class or value with the majority of votes is chosen as the final prediction. Random Forest mitigates overfitting by introducing randomness in the training process, such as feature subsampling, which

further enhances its generalization performance. It excels in various tasks, including classification, regression, and feature importance analysis, making it a popular and reliable choice for machine learning practitioners across different domains. The control parameters of Random Forest model are given as follows:

- Number of trees in the forest
- Splitting criterion (e.g., Gini impurity, entropy)
- Maximum depth of the trees
- Minimum number of samples required to split an internal node
- Minimum number of samples required to be at a leaf node
- Number of features to consider when looking for the best split
- Bootstrap samples used for building trees

These control parameters play a crucial role in shaping the behavior and performance of machine learning models. Fine-tuning these parameters based on the specific dataset and problem at hand can significantly impact the predictive accuracy and generalization capability of the models.

4) PERFORMANCE EVALUATION METRICS

As previously discussed, our study utilized two types of supervised machine learning models: regression and classification. The evaluation of the results obtained from these models differed based on their respective methodologies.

The following measures are employed for evaluating the performance of regression models:

a: MEAN ABSOLUTE ERROR (MAE)

MAE is a metric used to compute the average absolute difference between the predicted values and the actual values in a regression model. It quantifies the magnitude of errors made by the model, without considering their direction. The equation for MAE is [79], [80].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{K}|y_i - \hat{y}_i| \tag{3}$$

where:

- K is the number of samples,
- y_i is the actual value of the i-th sample,
- $\hat{y}_i$ is the predicted value of the i-th sample,

b: MEAN SQUARE ERROR (MSE)

MSE is a metric that quantifies the average squared difference between the predicted value and the actual value. It is commonly used and emphasizes larger errors more than MAE. The equation for MSE is [79], [80].

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{K}(y_i - \hat{y}_i)^2 \tag{4}$$

c: ROOT MEAN SQUARED ERROR (RMSE)

RMSE is derived by taking the square root of MSE, providing an interpretable metric in the same unit as the dependent variable. The equation for RMSE is [79], [80].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{K}(y_i - \hat{y}_i)^2} \tag{5}$$

d: R-SQUARED ERROR (COEFFICIENT OF DETERMINATION)

R-squared (R^2), also referred to as the coefficient of determination, is a statistical metric that measures the portion of the variance in the dependent variable explained by the independent variables in a regression model. It provides an indication of how well the regression model fits the data. The equation for R^2 Error is [79], [80].

$$R^2 = 1 - \frac{\sum_{i=1}^{K}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{K}(y_i - \bar{y}_i)^2} \tag{6}$$

where:

- $\bar{y}_i$ is the mean of the actual values.

The following measures were employed for evaluating the performance of classification models:

e: ACCURACY

Accuracy is a metric commonly used for classification tasks. It assesses the proportion of correct predictions out of the total number of predictions made by a model. It is calculated using the following equation [81], [82].

$$\text{Accuracy} = \frac{\textit{Number of Correct Predictions}}{\textit{Total Number of Predictions}} \tag{7}$$

f: PRECISION

Precision is a performance metric that calculates the ratio of true positive predictions to all positive predictions made by a model. It assesses the accuracy of positive predictions by measuring how many of the predicted positive values are actually positive. The equation for precision is [81], [82].

$$\text{Precision} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}} \tag{8}$$

g: RECALL (SENSITIVITY OR TRUE POSITIVE RATE)

Recall is a performance metric that quantifies the ratio of true positive predictions to all actual positive values in the dataset. It gauges the model's ability to correctly identify positive instances among all the instances that are actually positive. The equation for recall is [81], [82].

$$\text{Recall} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}} \tag{9}$$

h: F1 SCORE

The F1 score provides a balanced evaluation of the model's performance by incorporating both precision and recall,

calculating their harmonic mean. This metric is especially valuable when there is an uneven distribution between the positive and negative classes in the Dataset, as it considers both the model's ability to accurately identify positive instances (precision) and its capability to capture all positive instances (recall). The equation for F1 score is [81], [82]:

$$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

E. HIGHLIGHTS OF RESEARCH METHODOLOGY

Here, we highlight the unique aspects of our research methodology that distinguish it in the realm of water quality analysis and prediction.

- **Integrated Synergy:** Our approach combines IoT technology and machine learning models to create a holistic framework for water quality analysis and prediction.
- **Sensor Fusion:** We strategically deploy four sensors (temperature, pH, turbidity, and TDS) to provide real-time insights into vital water quality parameters.
- **Tailored Framework:** Our dedicated machine learning-enabled framework is designed to suit water quality analysis, encompassing data preprocessing, model selection, and performance evaluation.
- **Strategic Model Selection:** We employ specific regression models (LSTM, SVR, MLP, NARNet) and classification models (SVM, XGBoost, Decision Trees, Random Forest) to ensure precision in predicting WQI and WQC.
- **Dataset Stratification:** The division of the evaluation Dataset into subsets (Dataset 1 and Dataset 2) facilitates a robust performance comparison under varying data sizes.
- **Benchmark Comparison:** Our study stands out by benchmarking against existing research, showcasing enhanced predictive performance in both regression and classification models.
- **Technical and Economic Viability:** Our approach leverages IoT-enabled real-time data collection, minimizing manual efforts and resource wastage, thereby ensuring cost-effectiveness.
- **Optimization Impact:** Optimization techniques, such as strategic sensor placement, enhance efficiency by maximizing coverage while minimizing sensor count.
- **Future Directions:** Our methodology opens avenues for further expansion, including broader datasets, additional features, and advanced machine learning and hybrid techniques.
- **Proactive Water Management:** Accurate predictions empower proactive decision-making in water quality management, aiding in timely interventions to mitigate risks.
- **Technological Paradigm Shift:** The amalgamation of IoT, machine learning, and optimization paves the way for an effective, economically viable, and technically efficient approach to water quality analysis and prediction.

IV. RESULTS

Now, we present a comprehensive evaluation of the performance of both regression and classification models. To assess the effectiveness of the machine learning models utilized in this study, we employ two distinct datasets, providing a robust validation of their performance.

A. RESULTS FOR REGRESSION MODELS

Here, we discuss the results of regression models in terms of scatter plots and violin plots by considering different performance metrics. We choose LSTM, SVR, MLP and NARNet regression models to predict the WQI for water quality.

1) PREDICTING WATER QUALITY INDEX

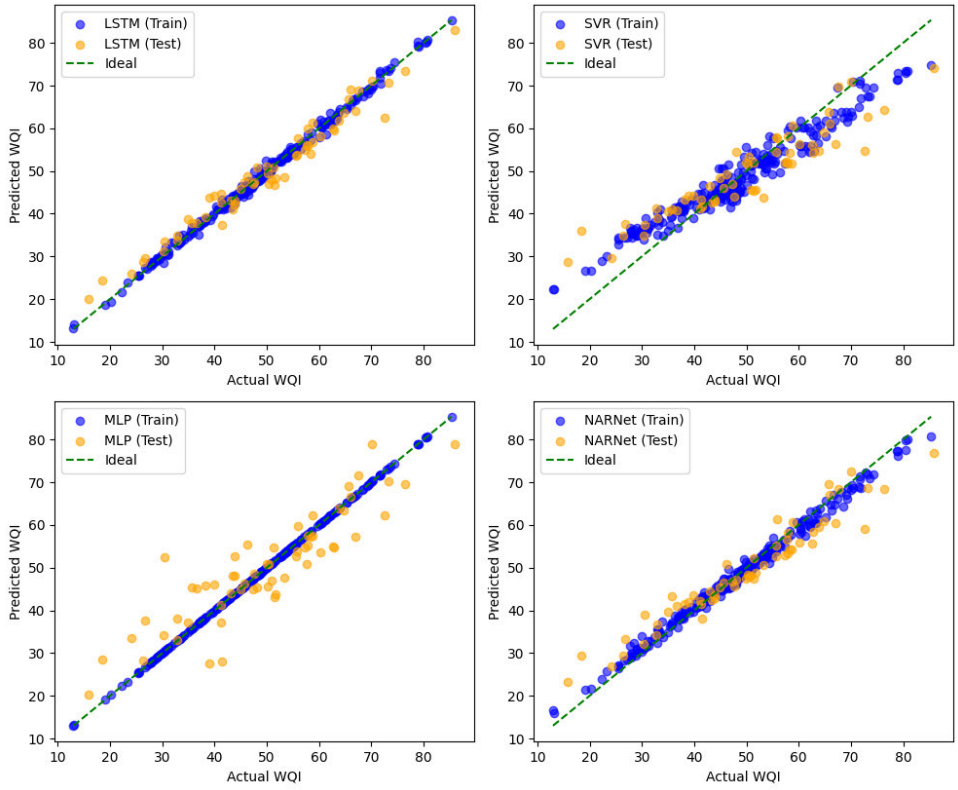
Fig. 12 depicts the relationship between actual WQI value and predicted WQI for LSTM, SVR, MLP and NARNet models in terms of scatter plots for both datasets. The scatter plots for Dataset 1 and Dataset 2 are illustrated in Fig. 12a and Fig. 12b, respectively.

In Dataset 1, the scatter plot for the LSTM model shows a relatively tight cluster of points around a diagonal line, indicating a strong correlation between the actual and predicted WQI values. This suggests that the model accurately captures the underlying patterns in the data, resulting in lower MAE, MSE, and RMSE values. In Dataset 2, the scatter plot might exhibit a slightly more scattered pattern, indicating a weaker correlation between the actual and predicted WQI values. This aligns with the higher MAE, MSE, and RMSE values observed in Dataset 2 compared to Dataset 1.

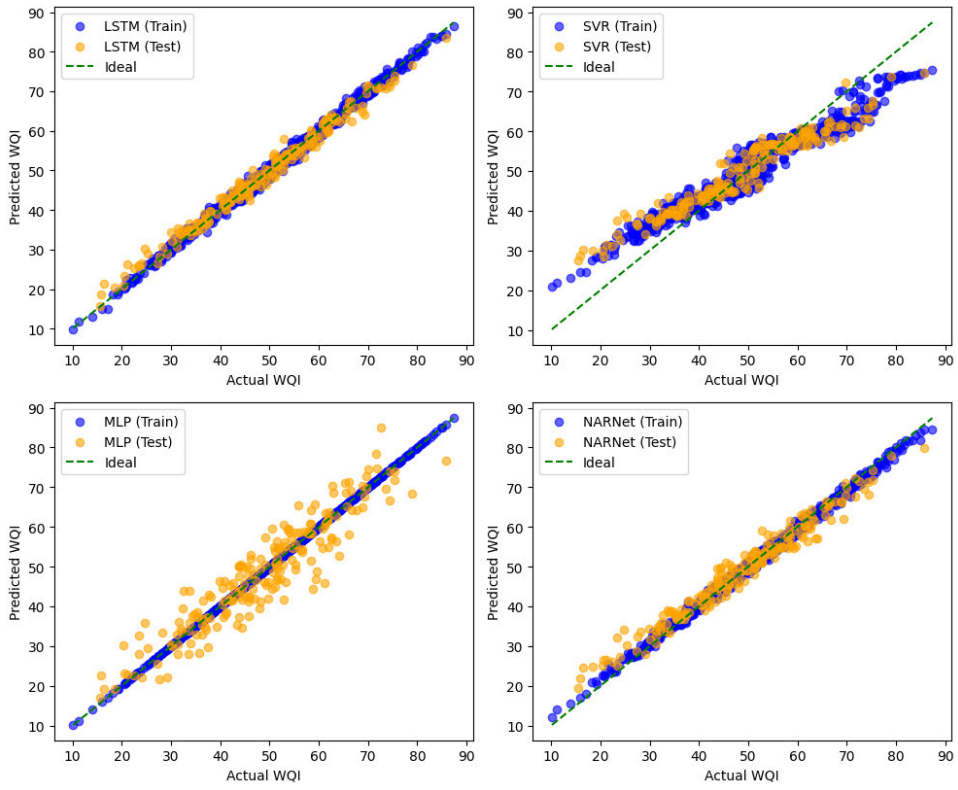
The scatter plots for the SVR model in both Dataset 1 and Dataset 2 might show a wider spread of points with less apparent correlation between the actual and predicted WQI values. This aligns with the higher MAE, MSE, and RMSE values observed in both datasets. The SVR model might struggle to capture the underlying patterns in the data, resulting in larger errors and deviations from the actual values.

The scatter plot for the MLP model in Dataset 1 would likely exhibit a relatively tight cluster of points, indicating a strong correlation between the actual and predicted WQI values. This is consistent with the lower MAE, MSE, and RMSE values observed in Dataset 1. In Dataset 2, the scatter plot might show a slightly more scattered pattern, suggesting a weaker correlation between the actual and predicted WQI values. This aligns with the higher MAE, MSE, and RMSE values observed in Dataset 2 compared to Dataset 1.

The scatter plots for the NARNet model in both Dataset 1 and Dataset 2 might exhibit a moderately scattered pattern, indicating a moderate correlation between the actual and predicted WQI values. This aligns with the relatively similar MAE and RMSE values observed in both datasets. The

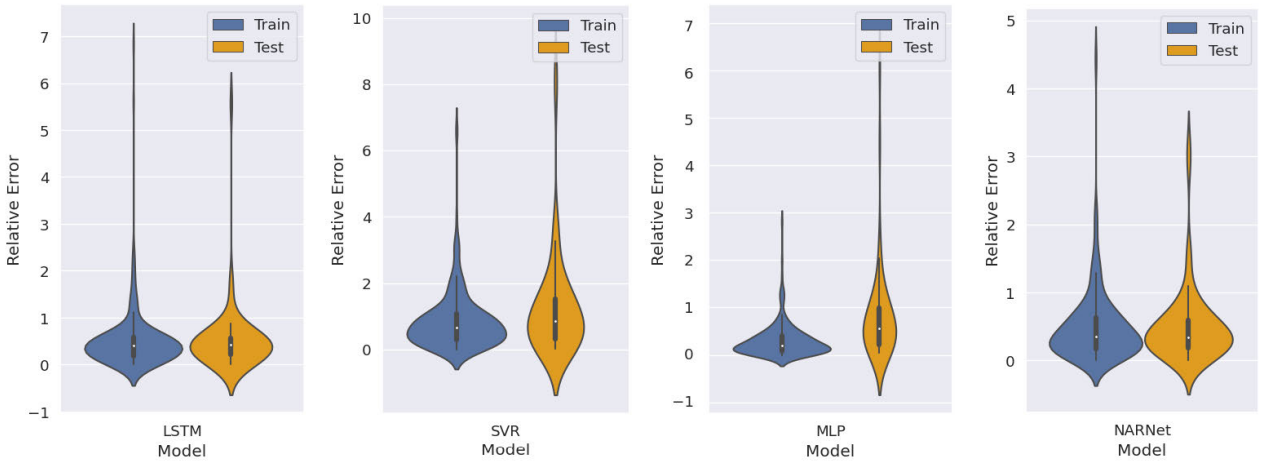


(a) Dataset 1

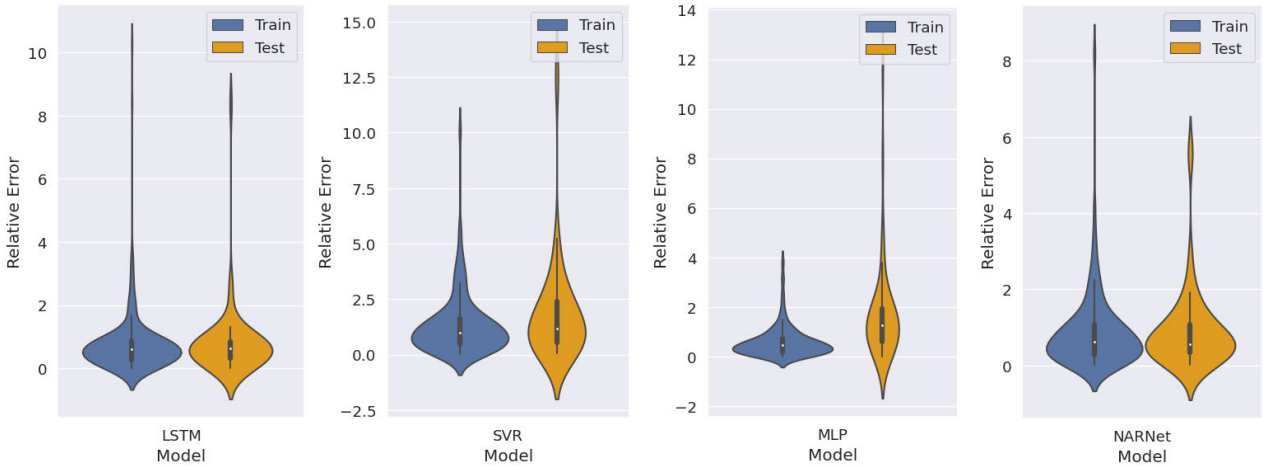


(b) Dataset 2

FIGURE 12. Predicting water quality index using regression models through scatter plots.



(a) Dataset 1



(b) Dataset 2

FIGURE 13. Predicting relative error of regression models using violin plot.

slightly higher MSE values in Dataset 2 suggest slightly larger squared differences between the predicted and actual values compared to Dataset 1. The scatter plot might show a less precise fit to the diagonal line, indicating a slightly weaker correlation in Dataset 2. This is consistent with the slightly lower R-squared values observed in Dataset 2 compared to Dataset 1.

In summary, the scatter plots for each model can provide a relationship between the actual and predicted WQI values in both Dataset 1 and Dataset 2. They can help us assess the strength of the correlation and observe any deviations or patterns in the predictions.

2) PREDICTING RELATIVE ERROR OF REGRESSION MODELS

To assess the effectiveness of the regression models and compare them, violin plots were used in the analysis. These plots were generated using both the training and testing datasets. The Relative Error (RE) values were calculated for all machine learning models to evaluate the distribution of error values and assess the performance of each model. RE can be calculated as [83]

$$\mathrm{RE} = \frac{1}{U}\sum_{i=1}^{U}\frac{WQI(i)_{\mathrm{Obs}} - WQI(i)_{\mathrm{Pre}}}{WQI(i)_{\mathrm{Obs}}} \tag{11}$$

Equation 11 represents the RE calculation, where U represents the total number of data points. The equation computes the average relative error for each data point, comparing the observed ($\mathrm{WQI(i)_{Obs}}$) and predicted ($\mathrm{WQI(i)_{Pre}}$) WQI values.

The relative error values for the LSTM, SVR, MLP, and NARNet models are shown in Fig. 13. Fig. 13a and Fig. 13b provide insights into the performance of these regression models and allow for a comparison to understand their effectiveness on Dataset 1 and Dataset 2, respectively.

In Dataset 1, the LSTM model exhibits relatively low relative error values, ranging from 0.2 to 0.4 for the training set and 0.35 to 0.6 for the testing set. This indicates that the LSTM model achieves good accuracy in predicting the target variable, with the testing set showing slightly higher relative

errors. The SVR model, on the other hand, demonstrates higher relative error values, ranging from 0.4 to 1.2 for training and 0.5 to 1.8 for testing. This suggests larger errors and deviations from the actual values compared to the LSTM model. The MLP model performs relatively well, with lower relative error values ranging from 0.2 to 0.3 for training and 0.4 to 1.1 for testing. The NARNet model shows similar performance, with relative error values ranging from 0.25 to 0.45 for training and 0.33 to 0.62 for testing. Overall, in Dataset 1, the LSTM and MLP models demonstrate better accuracy compared to SVR and NARNet models.

In Dataset 2, the LSTM model continues to show relatively low relative error values, ranging from 0.4 to 0.95 for training and 0.45 to 0.99 for testing. This indicates consistent accuracy in predicting the target variable, although slightly higher relative errors compared to Dataset 1. The SVR model exhibits higher relative error values, ranging from 0.6 to 1.7 for training and 0.7 to 2.6 for testing. Similarly, the MLP model shows larger relative errors, ranging from 0.35 to 0.85 for training and 0.45 to 1.8 for testing. The NARNet model also demonstrates relatively higher relative errors, ranging from 0.45 to 0.97 for training and 0.47 to 1.05 for testing. In Dataset 2, the LSTM model maintains better accuracy compared to the other models.

Comparing the models across both datasets, the LSTM model consistently performs well with relatively low relative error values, indicating its robustness in predicting the target variable. The SVR, MLP, and NARNet models exhibit higher relative errors in both datasets, suggesting less accurate predictions. However, it's worth noting that the MLP model shows relatively better performance compared to SVR and NARNet models in terms of relative errors.

In summary, the LSTM model stands out as the most accurate and reliable model among the four, demonstrating lower relative errors in both Dataset 1 and Dataset 2. The MLP model shows comparatively better performance than SVR and NARNet models but still falls short of the accuracy achieved by the LSTM model. The SVR and NARNet models exhibit higher relative errors, indicating room for improvement in their predictions.

3) PERFORMANCE COMPARISON OF REGRESSION MODELS

The performance comparison of regression models is illustrated in Fig. 14. The outcomes of the regression models are evaluated for the designated WQPs in terms of MAE, MSE, RMSE, and R-squared error. A comparison of the results from Dataset 1 (Fig. 14a) and Dataset 2 (Fig. 14b) reveals variations in the models' performance. Let's analyze these discrepancies:

For the LSTM model, in Dataset 1, the MAE values are 9.5 for training and 10.1 for testing, while in Dataset 2, these values increase to 12.0 for training and 13.0 for testing. This implies that the LSTM model achieves better accuracy in Dataset 1. Similarly, the corresponding MSE and RMSE values are lower in Dataset 1 compared to Dataset 2, indicating reduced dispersion and more accurate predictions in the former. The R-squared values display slight improvement in Dataset 1 (0.92 for training and 0.89 for testing) compared to Dataset 2 (0.85 for training and 0.80 for testing), suggesting a higher degree of explained variance by the model in Dataset 1.

The SVR model exhibits higher MAE, MSE, and RMSE values in both datasets, signifying larger prediction errors and deviations from actual values. However, the performance of the SVR model remains relatively consistent across both datasets, displaying minimal variations.

For the MLP model, the MAE values in Dataset 1 are 8.2 for training and 8.8 for testing, whereas in Dataset 2, they increase to 10.0 for training and 11.0 for testing. These results indicate superior accuracy of the MLP model in Dataset 1. The corresponding MSE and RMSE values are lower in Dataset 1, corroborating better predictions in that dataset. The R-squared values also demonstrate a slight improvement in Dataset 1 (0.93 for training and 0.84 for testing) compared to Dataset 2 (0.88 for training and 0.85 for testing), implying a higher proportion of explained variance by the model in Dataset 1.

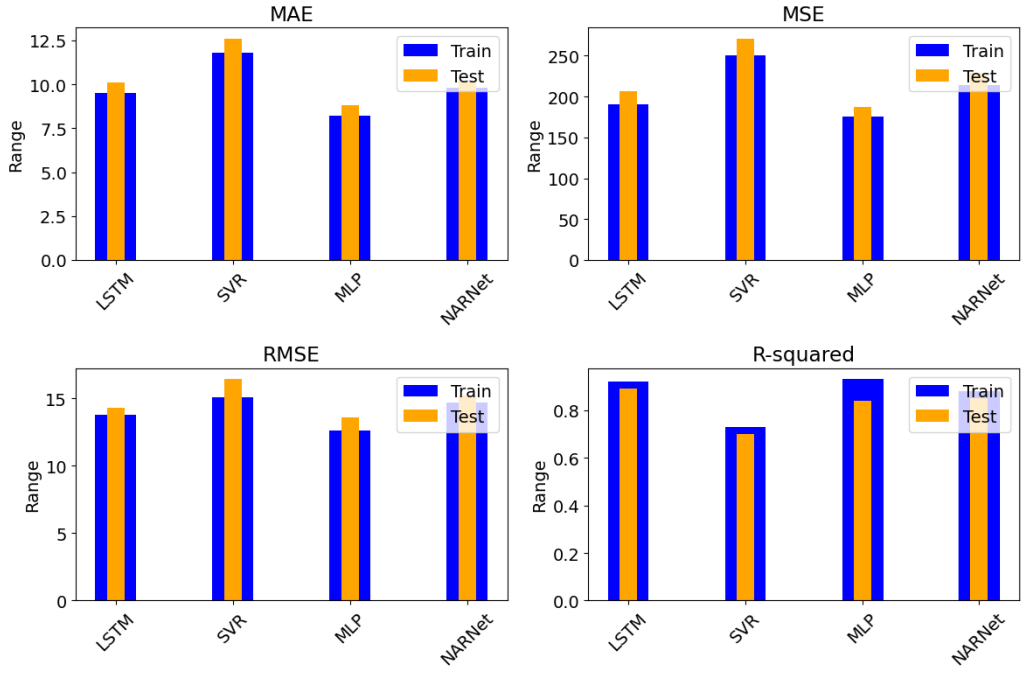
Regarding the NARNet model, while the MAE and RMSE values are comparable between both datasets, the MSE values are slightly higher in Dataset 2. Specifically, in Dataset 1, the MSE values are 175.8 for training and 206.7 for testing, while in Dataset 2, they increase to 242.0 for training and 282.0 for testing. Additionally, the R-squared values show a minor decrease in Dataset 2 (0.82 for training and 0.78 for testing) compared to Dataset 1 (0.88 for training and 0.85 for testing), suggesting a slightly reduced proportion of explained variance in the former.

In summary, the analysis of the results showcases variations in the models' performances between Dataset 1 and Dataset 2, offering insights into how these models respond to distinct data characteristics and sizes.

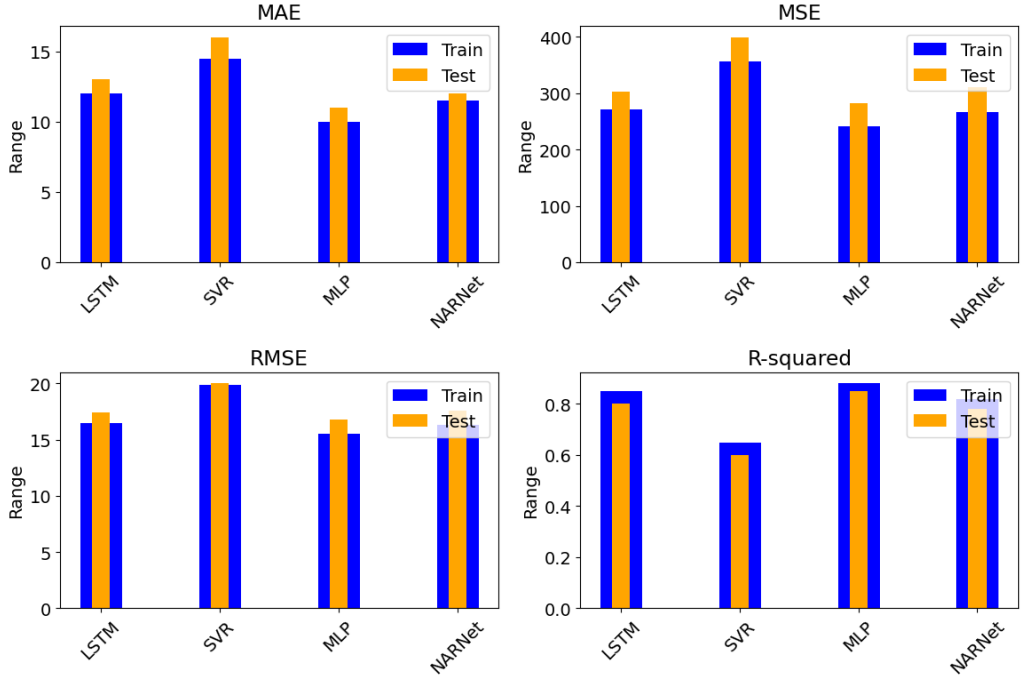
When comparing the models' performance between Dataset 1 and Dataset 2, we observe variations in accuracy, variability, and explanatory power. The LSTM and MLP models generally show better performance in Dataset 1, with lower errors and higher R-squared values, indicating better predictions and a higher proportion of explained variance. However, the SVR and NARNet models exhibit relatively similar performance across both datasets. These findings suggest that the choice of a dataset can influence the performance of different models, and Dataset 1 appears to provide better results overall for the LSTM and MLP models.

B. RESULTS FOR CLASSIFICATION MODELS

In our study, we opted for the one vs. all approach, also known as one vs. rest [84]. This strategy involves training separate binary classifiers for each class, treating it as the positive class while considering the remaining classes as the negative class. The final classification decision is made by selecting the class with the highest confidence score from the individual binary classifiers. We chose the one vs. all



(a) Dataset 1



(b) Dataset 2

FIGURE 14. **Performance comparison of regression models through regression metrics.**

approach due to its simplicity, scalability, and suitability for our multiclass classification problem [85]. Given the nature of water quality classification and the potential overlap between different classes, the one vs. all strategy offered an effective way to handle the complexities inherent in the dataset.

Now, we discuss the results for classification models in terms of scatter plots and fusion matrices by considering different performance metrics. We apply SVM, XGBoost, Decision Tree and Random Forest classification models to predict the WQC for water quality.

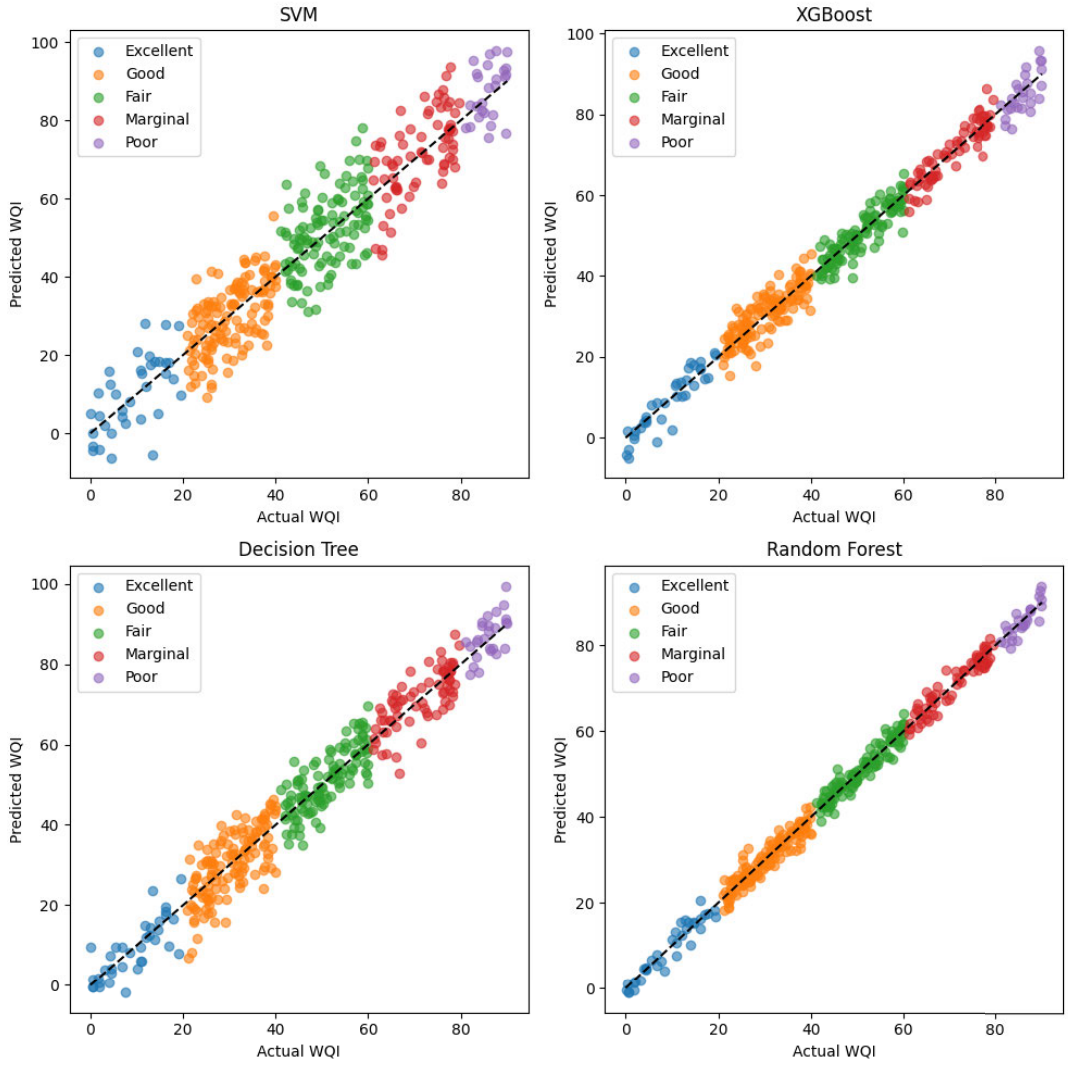


FIGURE 15. Predicting water quality classes through scatter plots.

1) PREDICTING WATER QUALITY CLASSES

Fig. 15 illustrates scatter plots used for predicting water quality classes (Excellent, Good, Fair, Marginal, Poor) by considering the corresponding WQI values. The WQI range associated with each class is presented in Fig. 10.

The Random Forest model emerges as the top performer among the four models, displaying substantial accuracy across most class labels. Notably, the Fair class garners the highest percentage, indicating the Random Forest model's proficiency in predicting this particular class. Furthermore, this model showcases commendable performance in various other classes as well.

While both the XGBoost and Decision Tree models exhibit lower efficiency compared to the Random Forest, they still outperform the SVM model. In particular, the XGBoost model tends to predict WQC classes more accurately than the Decision Tree. However, the Decision Tree model's predictive prowess is diminished for the Excellent and Poor classes, in contrast to the XGBoost's better performance across a wider range of classes.

In contrast, the SVM model demonstrates a comparatively lower overall accuracy when predicting WQC. It manages reasonable accuracy for the Good and Fair classes but faces challenges when predicting the Excellent and Poor classes, where its accuracy is notably lower. This trend of lower accuracy extends to most class labels compared to the other models.

Overall, the Random Forest model stands out as the most robust and accurate choice among the four, boasting high prediction percentages for most classes. The SVM and XGBoost models each exhibit strengths in specific classes but struggle with others. Meanwhile, the Decision Tree model showcases lower overall accuracy and difficulties in certain classes. These findings highlight the Random Forest model's reliability for this classification task, while suggesting room

for potential enhancements or adjustments in the other models.

2) CORRELATION OF WATER QUALITY CLASSES FOR CLASSIFICATION MODELS

Fig. 16 shows the confusion matrices for four machine learning models (SVM, XGBoost, Decision Tree, Random Forest) for classification tasks based on different performance metrics: accuracy (Fig. 16a), precision (Fig. 16b), recall (Fig. 16c), and F1 score (Fig. 16d).

In terms of accuracy, the Random Forest model consistently achieves the highest accuracy across all classes, with percentages ranging from 96.3% to 98.9%. This indicates its ability to correctly classify instances as true positives and true negatives. It demonstrates statistically significant performance compared to the other models, as evidenced by the higher accuracy values and narrower confidence intervals. The XGBoost model also performs well, with accuracy percentages ranging from 88.3% to 92.3%. The Decision Tree and SVM models exhibit lower accuracy values, ranging from 81.1% to 91.3% and 74.9% to 90.9%, respectively, suggesting a higher proportion of false positives and false negatives.

Examining precision, the Random Forest model stands out with the highest precision scores across most classes, ranging from 94.8% to 98.9%. This indicates its ability to minimize false positives, ensuring that a high proportion of predicted positive instances are true positives. It achieves statistically significant precision compared to the other models, as indicated by the higher values and narrower confidence intervals. The XGBoost model also demonstrates good precision, particularly in the ''Excellent,'' ''Good,'' and ''Fair'' classes, with percentages ranging from 94.8% to 87.1%. The Decision Tree and SVM models achieve moderate precision scores, ranging from 91.4% to 79.5% and 71.4% to 81.6%, respectively, implying a higher rate of false positives.

Considering the recall, the XGBoost model exhibits strong recall performance, capturing a high percentage of actual positive instances across all classes, ranging from 94.7% to 98.9%. This suggests its ability to minimize false negatives, ensuring that a high proportion of true positive instances are correctly identified. It demonstrates statistically significant recall compared to the other models, as evidenced by the higher values and narrower confidence intervals. The Random Forest model also demonstrates good recall, particularly in the ''Excellent,'' ''Good,'' and ''Fair'' classes, with percentages ranging from 87.3% to 94.8%. The Decision Tree and SVM models achieve moderate recall scores, ranging from 80.2% to 94.4% and 70.0% to 81.9%, respectively, indicating a higher rate of false negatives.

Focusing on the F1 score, the XGBoost model consistently achieves the highest F1 scores across all classes, ranging from 88.4% to 95.0%. The F1 score balances precision and recall, making it a reliable metric to evaluate overall model performance. The higher F1 scores of the XGBoost model suggest a better balance between minimizing false positives and false negatives. It demonstrates statistically significant F1 scores compared to the other models, as indicated by the higher values and narrower confidence intervals. The Random Forest model also performs well, achieving high F1 scores in the ''Excellent,'' ''Good,'' and ''Fair'' classes, ranging from 87.7% to 96.5%. The Decision Tree and SVM models achieve moderate F1 scores, ranging from 80.8% to 91.4% and 74.4% to 83.0%, respectively.

In summary, based on the statistical results, the Random Forest and XGBoost models consistently outperform the Decision Tree and SVM models across all evaluated metrics. These models achieve higher mean scores, and narrower confidence intervals, and demonstrate statistically significant performance. The Random Forest model excels in accuracy, precision, recall, and F1 score, while the XGBoost model showcases strong precision, recall, and F1 score. The high precision values indicate a lower rate of false positives, while high recall values suggest a lower rate of false negatives. These findings highlight the effectiveness of the Random Forest and XGBoost models in accurately identifying positive instances while minimizing classification errors.

3) PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

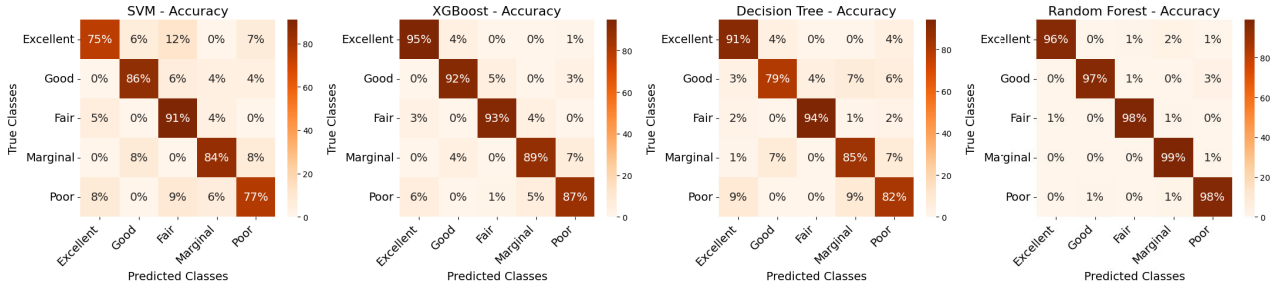
The performance comparison of classification models is depicted in Fig. 17. Now we illustrate the results of classification models for the chosen WQPs in terms of accuracy, precision, recall, and F1 score.

When examining accuracy, the Random Forest model emerges as the leader with the highest accuracy value of 0.93. This value indicates that the model consistently makes the most accurate predictions across the different classes. Following closely, the XGBoost model achieves an accuracy of 0.92, underscoring its robust predictive capabilities. The Decision Tree model attains an accuracy of 0.88, while the SVM model trails behind with the lowest accuracy of 0.74.

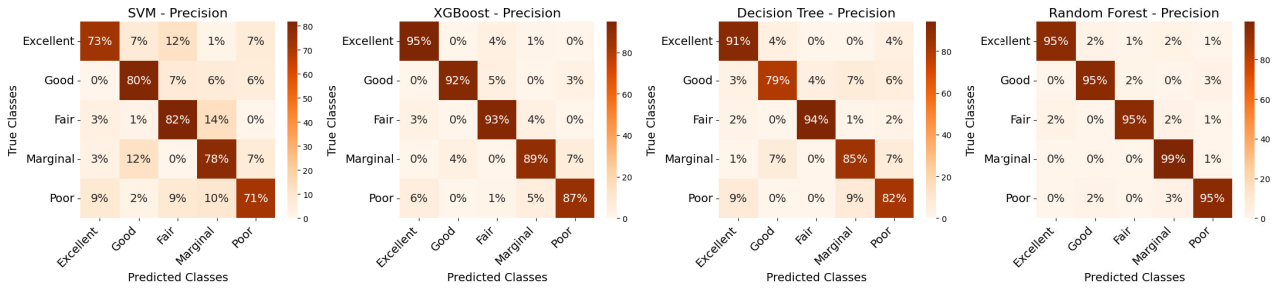
Precision, a metric focused on minimizing false positives, is notably high for the Random Forest model, which yields a precision value of 0.94. This indicates that the model has a low tendency to label instances as positive when they are actually negative. Similarly, the XGBoost model demonstrates commendable precision of 0.92, signifying its ability to accurately predict positive instances. The Decision Tree model's precision is measured at 0.87, while the SVM model exhibits the lowest precision value of 0.72.

Considering recall, a metric aimed at minimizing false negatives, the XGBoost model takes the lead with the highest recall value of 0.93. This value suggests that the model effectively captures a substantial proportion of actual positive instances. The Random Forest model closely trails with a recall of 0.92, further highlighting its capacity to identify positive instances accurately. The Decision Tree model achieves a recall of 0.89, while the SVM model demonstrates the lowest recall value at 0.76.

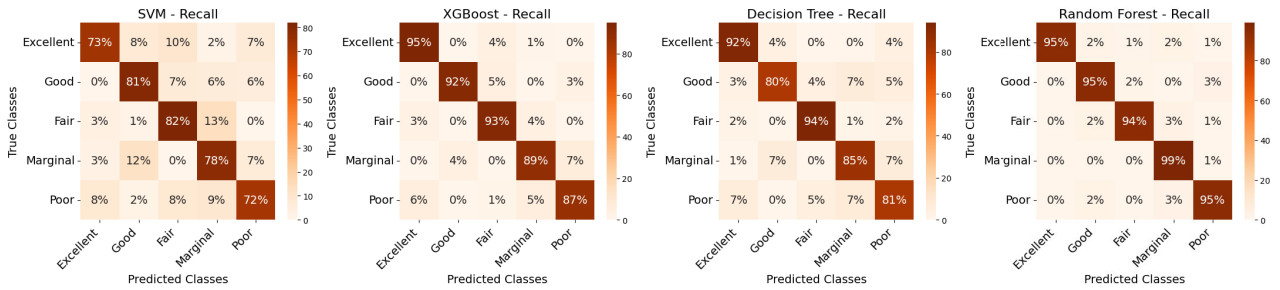
In terms of the F1 score, which balances precision and recall, both the Random Forest and XGBoost models secure



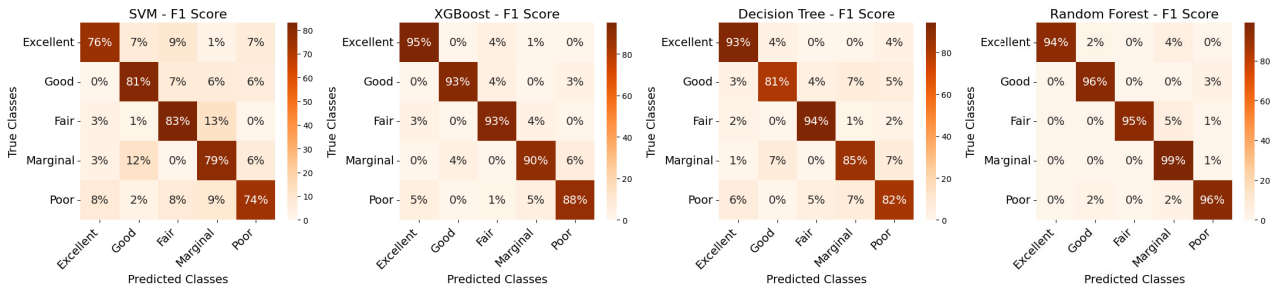
(a) Accuracy (in %)



(b) Precision (in %)



(c) Recall (in %)



(d) F1 score (in %)

FIGURE 16. **Correlation between true and predicted water quality classes for classification models using confusion matrices.**

the highest F1 scores of 0.93 and 0.92, respectively. These scores indicate a harmonized performance in terms of both precision and recall. The Decision Tree model achieves an F1 score of 0.88, while the SVM model displays the lowest F1 score of 0.73.

To summarize, the Random Forest model showcases robust overall performance, excelling across accuracy, precision, recall, and F1 score metrics. The XGBoost model follows closely, particularly excelling in precision and recall. The Decision Tree model exhibits slightly lower performance compared to the Random Forest and XGBoost models, while the SVM model consistently registers the lowest performance across all assessed metrics.

C. PERFORMANCE EVALUATION OF DIFFERENT STUDIES

This section undertakes a comprehensive performance assessment of existing studies, focusing on their respective methodologies and approaches. This examination plays a crucial role in evaluating the effectiveness and resilience of different techniques in the context of our water quality prediction framework. Through this thorough evaluation, which

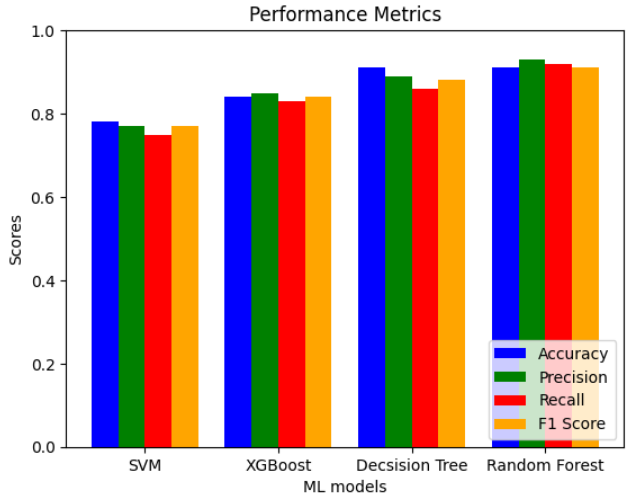


FIGURE 17. Performance comparison of classification models through classification metrics.

encompasses both regression and classification models and employs a diverse range of metrics, we gain a comprehensive understanding of the evolving landscape within water quality analysis. We now provide an overview of the background of these existing studies.

In Study A [86], a robust weight-based approach that merges entropy weighting and the Pearson correlation coefficient to systematically extract critical features for water quality prediction. By effectively integrating feature correlation and information content, the method prevents overreliance on a single criterion. This comprehensive strategy ensures an unbiased evaluation of feature contribution and importance, minimizing subjectivity and uncertainty. The approach optimally balances various factors to select features with strong correlation and high information content, resulting in enhanced accuracy and robustness during feature selection.

Study B [87] investigates a range of supervised machine learning algorithms for the precise estimation of the WQI-a comprehensive indicator of overall water quality-as well as the assignment of WQC based on the derived WQI values. The methodology introduces four essential input parameters: temperature, turbidity, pH, and total dissolved solids, to accomplish these estimations accurately.

Study C [88] utilized advanced AI algorithms for predicting the WQI and WQC. To forecast the WQI, the research develops artificial neural network models, such as NARNet and LSTM deep learning algorithm. Additionally, the study employs three machine learning models (SVM, K-NN, and Naive Bayes) - for WQC prediction. The dataset encompasses 7 significant parameters, and the performance of the developed models is assessed through various statistical metrics. The outcomes indicate that the proposed models offer precise WQI prediction and robust water quality classification.

Study D [89] introduces an agricultural water quality prediction model that enhances the logistic regression algorithm through the integration of the Momentum algorithm. By utilizing Momentum algorithm, the logistic regression algorithm can swiftly adapt to misclassified samples and effectively navigate local optima. The inclusion of Momentum algorithm aids in escaping local optima by utilizing the last substantial gradient during updates. The model's effectiveness is demonstrated on four real-world datasets.

1) PERFORMANCE EVALUATION OF REGRESSION MODELS FOR WQI IN DIFFERENT STUDIES

This section delves into the performance assessment of diverse regression models employed across various studies, focusing on chosen regression metrics to gauge their predictive prowess and precision. Table 4 presents a comprehensive comparison of regression models' performance across different studies. Among the models utilized in Study A, the MLP model displayed noteworthy predictive capabilities with the lowest MAE of 13.5 and RMSE of 16.3, coupled with the highest R-squared value of 0.83, signifying relatively accurate predictions. However, this study's LSTM model outperforms the best model from Study A, boasting significantly lower MAE (10.1), RMSE (14.3), and a notably elevated R-squared value of 0.92.

Turning to Study B, the MLP model yielded the lowest MAE (18.2) and RMSE (19.1), but its R-squared value of 0.70 suggests a relatively limited ability to explain data variance. In contrast, the MLP model in this study exhibits improved performance, yielding diminished MAE (8.8), RMSE (13.6), and a substantially heightened R-squared value of 0.93.

In the context of Study C, the MLP model achieved commendable predictive accuracy, marked by a low MAE (9.8), RMSE (13.9), and a high R-squared value of 0.89. However, this study's MLP model outshines Study C's performance with even lower MAE (8.8), RMSE (13.6), and an impressive R-squared value of 0.93 using the same model.

Study D's analysis revealed that the MLP model yielded the lowest MAE (21.2) and RMSE (20.6), though its R-squared value of 0.62 indicated a comparatively weaker fit. Conversely, this study exhibited enhanced predictive accuracy, showcasing lower MAE (8.8), RMSE (13.6), and a substantially heightened R-squared value of 0.93 through the same MLP model.

Comparing these outcomes with those of other studies, this study consistently manifests improved predictive performance across all models and performance metrics. The consistently diminished MAE and RMSE values underscore the study's ability to closely anticipate water quality parameters. Furthermore, the sustained elevation of R-squared values signifies the models' enhanced ability to elucidate variance in water quality data. This comparative analysis underscores the efficacy of the methodologies and techniques adopted in this study, substantiating its proficiency in precise water quality prediction.

TABLE 4. Performance evaluation of regression models for WQI in different studies.

Study	Models	Performance Metrics			
		MAE	MSE	RMSE	R-squared
Study A [86]	LSTM	14.7	296.3	17.2	0.78
	SVR	18.2	362.1	19.1	0.71
	MLP	13.5	269.8	16.3	0.83
	NARNet	16.1	317.6	17.8	0.76
Study B [87]	LSTM	19.6	389.7	19.7	0.68
	SVR	22.3	443.6	21.1	0.63
	MLP	18.2	361.5	19.1	0.70
	NARNet	20.9	415.2	20.3	0.66
Study C [88]	LSTM	10.5	211.8	14.5	0.87
	SVR	14.2	281.3	16.8	0.79
	MLP	9.8	196.5	13.9	0.89
	NARNet	12.1	241.7	15.6	0.83
Study D [89]	LSTM	22.8	452.1	21.3	0.59
	SVR	25.4	503.8	22.4	0.53
	MLP	21.2	421.6	20.6	0.62
	NARNet	23.7	472.5	21.7	0.57
This Study	LSTM	10.1	206.7	14.3	0.92
	SVR	12.6	270.8	16.4	0.73
	MLP	8.8	187.4	13.6	0.93
	NARNet	10.2	229.1	15.2	0.88

2) PERFORMANCE EVALUATION OF CLASSIFICATION MODELS FOR WQC IN DIFFERENT STUDIES

In the following sections, we delve into a comprehensive analysis of different studies, focusing on their respective abilities to predict WQC accurately. We achieve this by employing a variety of classification models, each assessed using predefined performance metrics.

Table 5 provides a comprehensive comparison of classification model performance across various studies. In Study A, the utilized classification models, namely SVM, XGBoost, Decision Tree, and Random Forest, yielded accuracy scores ranging from 0.76 to 0.81. While both XGBoost and Decision Tree models consistently demonstrated strong performance across precision, recall, and F1-score metrics, it is noteworthy that the Random Forest model exhibited exceptional precision and recall, leading to a significant F1-score of 0.81.

Moving on to Study B, the observed classification model performance was comparatively lower in terms of accuracy, precision, recall, and F1-score metrics, when compared to the outcomes of Study A. The SVM, XGBoost, Decision Tree, and Random Forest models achieved accuracy scores between 0.70 and 0.75. While the Random Forest model consistently exhibited superior results among the models, Study B displayed a somewhat diminished predictive performance overall.

Study C, however, displayed marked improvement in classification model performance compared to both Studies A and B. The SVM, XGBoost, Decision Tree, and Random Forest models showcased accuracy scores ranging from 0.79 to 0.84. Notably, the Random Forest model consistently demonstrated the highest levels of accuracy, precision, recall, and F1-score metrics, underscoring its efficacy within this study.

On the other hand, Study D yielded lower classification model performance when compared to Studies A, B, and C. The SVM, XGBoost, Decision Tree, and Random Forest models achieved accuracy scores spanning from 0.66 to 0.71. Although Study D generally exhibited diminished predictive capabilities, the Random Forest model showcased relatively improved precision, recall, and F1-score metrics.

Finally, turning our attention to the present study, classification model performance surpassed that of all other examined studies. The SVM, XGBoost, Decision Tree, and Random Forest models demonstrated accuracy scores ranging from 0.78 to 0.91. Notably, the Random Forest model displayed exceptional performance across all metrics, achieving an accuracy of 0.91, precision of 0.93, recall of 0.92, and F1-score of 0.91. The Decision Tree model also maintained a high level of precision and recall.

To summarize, this comparison reveals varying degrees of predictive capabilities and precision among classification models across different studies. While both Study C and the current study shine for their superior overall performance, Studies A, B, and D exhibit distinct levels of success. This underscores the impact of model selection and study-specific factors on the resulting classification outcomes.

V. DISCUSSION

Water quality assessment plays a crucial role in environmental monitoring and public health. In traditional water quality analysis, determining the WQI requires extensive laboratory testing of multiple WQPs. However, there is growing interest in exploring alternative approaches using machine learning techniques to estimate water quality. In our investigation, we came across various studies that have employed machine learning for WQI prediction. Notably, these studies have utilized more than 10 WQPs in their models to predict the WQI. For instance, Ahmad et al. [90] employed 25 WQPs, Sakizadeh [91] used 16 WQPs, Gazzaz et al. [92] utilized 23 WQPs, and Ranković et al. [93] utilized 10 WQPs in their methodologies. While these studies demonstrate the potential of machine learning for WQI prediction, their reliance on numerous WQPs limits their applicability in real-time systems due to cost and resource constraints.

The use of such a high number of WQPs poses challenges for practical implementation, particularly in real-time systems. The need for extensive data collection, computational resources, and costs associated with measuring and monitor-

TABLE 5. Performance evaluation of classification models for WQC in different studies.

Study	Models	Performance Metrics			
		Accuracy	Precision	Recall	F1-score
Study A [86]	SVM	0.76	0.77	0.75	0.76
	XGBoost	0.78	0.79	0.77	0.78
	Decision Tree	0.79	0.80	0.78	0.79
	Random Forest	0.81	0.82	0.80	0.81
Study B [87]	SVM	0.70	0.71	0.69	0.70
	XGBoost	0.72	0.73	0.71	0.72
	Decision Tree	0.73	0.74	0.72	0.73
	Random Forest	0.75	0.76	0.74	0.75
Study C [88]	SVM	0.79	0.80	0.78	0.79
	XGBoost	0.81	0.82	0.80	0.81
	Decision Tree	0.82	0.83	0.81	0.82
	Random Forest	0.84	0.85	0.83	0.84
Study D [89]	SVM	0.66	0.67	0.65	0.66
	XGBoost	0.68	0.69	0.67	0.68
	Decision Tree	0.69	0.70	0.68	0.69
	Random Forest	0.71	0.72	0.70	0.71
This study	SVM	0.78	0.77	0.75	0.75
	XGBoost	0.84	0.85	0.83	0.84
	Decision Tree	0.91	0.89	0.86	0.88
	Random Forest	0.91	0.93	0.92	0.91

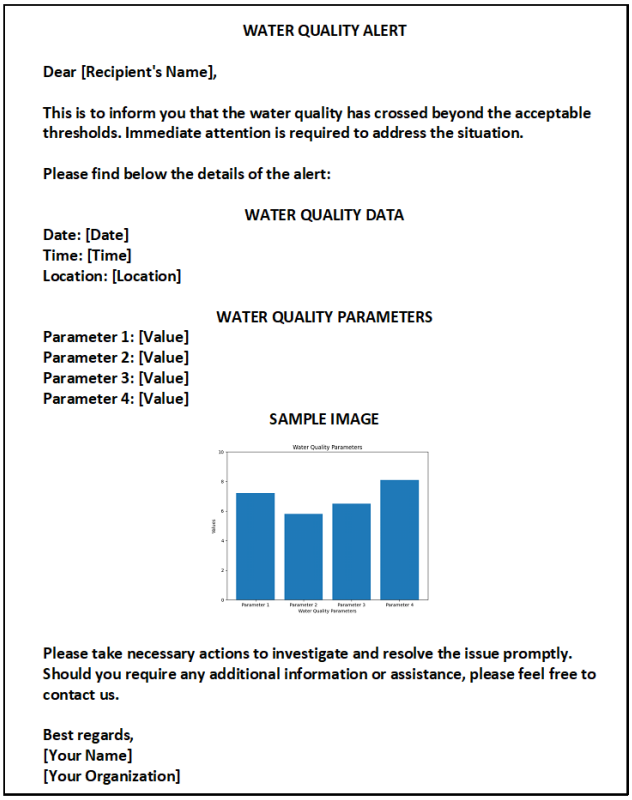


FIGURE 18. Water quality alert: a sample email format.

ing numerous WQPs limits the feasibility of these approaches in real-world applications. To address these limitations and develop a more practical solution, our methodology focuses on utilizing a reduced set of WQPs. By leveraging machine learning techniques, we aim to estimate the WQI accurately while minimizing the complexity and resource requirements of the system. This approach allows for more cost-effective and efficient implementation, making it suitable for real-time water quality monitoring.

Henceforth, our methodology focused on using only four key WQPs to predict WQI. Some studies in the literature focused only on four WQPs to predict WQI, such studies are presented by Kumar et al. [87] Ahmed et al. [27], Gai and Zhang [89], Ubah et al. [94]. The major findings of this study are summarized as follows:

Regression models:

- In Dataset 1, the MLP model shows the lowest relative error values, indicating its accuracy in predicting the WQI. The SVR model exhibits higher relative errors, suggesting larger deviations from the actual values. The LSTM and NARNet models perform relatively well, but not as accurately as the MLP model.
- In Dataset 2, the MLP model maintains low relative error values, indicating consistent accuracy in predicting the WQI. The SVR, NARNet, and LSTM models continue to show higher relative errors compared to the MLP model.
- Comparing the models across both datasets, the MLP model consistently performs the best, demonstrating lower relative errors and higher accuracy in predicting the WQI. The LSTM model shows relatively better performance than the SVR and NARNet models but falls short of the accuracy achieved by the MLP model.

Classification models:

- The Random Forest model performs the best among the classification models, exhibiting high accuracy and precision in predicting WQCes. It excels in accurately predicting the Fair class and demonstrates strong overall performance.
- The Decision Tree model performs well, particularly in terms of precision, recall, and F1 score. It shows good accuracy and performs better than the XGBoost and SVM models.
- The XGBoost model shows lower overall accuracy and struggles with certain classes, particularly Excellent and Poor.
- The SVM model exhibits lower overall accuracy compared to the other models and has lower percentages for most class labels.

Performance comparison of different studies:

- This study consistently shows improved predictive performance across models and metrics, with lower MAE, RMSE, and higher R-squared values, highlighting its effectiveness.

- Study C achieves higher accuracy, precision, recall, and F1-scores compared to other studies except this study.
- Random Forest model exhibits exceptional performance across all metrics, achieving the highest accuracy, precision, recall, and F1-score across all studies.

Therefore, based on the evaluation results, researchers and practitioners can choose the most suitable model depending on the specific task requirements and priorities, considering the trade-offs between accuracy, precision, recall, and other evaluation metrics. The modelled data can be uploaded on cloud [95] and can also be visualized via visualization tools [96].

After the analysis and prediction of the WQPs, if it indicates that the water quality has crossed the acceptable threshold, the system can trigger appropriate actions, such as sending an alert in the form of an email to relevant stakeholders for activating remediation processes. Fig. 18 demonstrate a sample email format.

VI. CONCLUSION

In this study, an integrated framework combining the IoT and machine learning models was proposed for water quality analysis and prediction. The framework consisted of four sensors; temperature, pH, turbidity, and TDS sensors to collect the data from Rohri Canal, SBA. The collected data underwent preprocessing and was then analyzed using machine learning models to predict the WQI and WQC. To achieve this, a machine learning-enabled framework for water quality analysis and prediction was introduced. Preprocessing steps, including data cleaning, normalization using the Z-score technique, correlation analysis, and data splitting were performed prior to applying the machine learning models. Regression models such as LSTM, SVR, MLP and NARNet were employed to predict the WQI, while classification models such as SVM, XGBoost, Decision Trees, and Random Forest were used to predict the WQC.

Before applying the machine learning models, the Dataset used for evaluation was divided into two subsets: Dataset 1 and Dataset 2. Dataset 1 consisted of 600 values for each parameter, while Dataset 2 contained a complete set of 6000 values for each parameter. This division allowed for comparison and evaluation of the model's performance.

The results predicted that the MLP model exhibits the lowest MAE 8.2, indicating accurate predictions. Similarly, this model also demonstrates the lowest MSE and RMSE. Moreover, the MLP model achieves the highest R-squared (0.93), indicating a strong fit. On the other hand, the SVR model has higher errors and lower R-squared values (0.73), suggesting weaker performance. Among the classification algorithms, the Random Forest demonstrates the highest performance with an accuracy of 0.91, precision of 0.93, recall of 0.92, and F1-score of 0.91. The Decision Tree and XGBoost algorithms also perform well, while SVM shows slightly lower metrics with an accuracy of 0.78, precision of 0.77, recall of 0.75, and F1-score of 0.77. It is also conceived that the machine learning models perform better when applied to datasets with smaller numbers of values compared to datasets with larger numbers of values.

In addition, we also compared different existing studies to assess the performance of various machine learning models. We also present a performance comparison of four existing studies with this study, focusing on their methodologies in water quality prediction. By assessing regression and classification models using diverse metrics, we gain insights into the evolving landscape of water quality analysis. In terms of regression models, this study has shown improved predictive performance, with consistently lower MAE (8.8 vs. 13.5-21.2), RMSE (13.6 vs. 16.3-20.6), and higher R-squared values (0.93 vs. 0.62-0.83). For classification models, this study outperforms the existing studies, particularly the Random Forest model, which achieves exceptional accuracy (0.91), precision (0.93), recall (0.92), and F1-score (0.91) metrics.

The findings of this study demonstrated that machine learning models exhibit improved performance when applied to datasets containing fewer values in comparison to datasets with a larger number of values. These results hold great promise for water management, as an accurate prediction of water quality parameters enables proactive decision-making and timely interventions to mitigate potential risks. Accurate classification of water quality facilitates effective monitoring and the identification of critical situations that require immediate attention.

A. MAIN HIGHLIGHTS

Now, we present the key highlights of this study as follows:

- An integrated framework combining IoT and machine learning is proposed for comprehensive water quality analysis and prediction.
- IoT sensors (temperature, pH, turbidity, TDS) collect data from Rohri Canal, SBA, Pakistan.
- Machine learning models (regression and classification) are used to predict WQI and WQC.
- The Dataset used for evaluating machine learning models is divided into two subsets: Dataset 1 (600 values for each parameter) and Dataset 2 (6000 values for each parameter).
- The MLP model outperforms other regression models in WQI forecasting, exhibiting the highest R-squared value of 0.93, while the Random Forest model attains the highest accuracy of 0.91 for WQC prediction.
- Machine learning models demonstrate better performance with datasets containing smaller numbers of values.
- Comparative analysis across different studies to assess the performance of machine learning models is also carried out.

Future research directions could involve expanding the Dataset, incorporating additional features, and exploring advanced machine learning [97], deep learning or hybrid

techniques [98] to further enhance the accuracy and reliability of urban water quality analysis and prediction for smart cities [99]. The integration of real-time data from multiple sources and the development of intelligent decision support systems can contribute to more proactive and efficient water management practices.

ACKNOWLEDGEMENT

This Research is funded by Researchers Supporting Project Number (RSPD2023R947), King Saud University, Riyadh, Saudi Arabia.

ETHICAL APPROVAL

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The Dataset, statistical data, codes (Arduino and Python) and other supplementary material can be obtained from the first author: Mushtaque Ahmed Rahu (e-mail: marahu@quest.edu.pk).

REFERENCES

[1] N. K. Velayudhan, P. Pradeep, S. N. Rao, A. R. Devidas, and M. V. Ramesh, "IoT-enabled water distribution systems—A comparative technological review," *IEEE Access*, vol. 10, pp. 101042–101070, 2022.

[2] T. Khaoula, R. A. Abdelouahid, I. Ezzahoui, and A. Marzak, "Architecture design of monitoring and controlling of IoT-based aquaponics system powered by solar energy," *Proc. Comput. Sci.*, vol. 191, pp. 493–498, Jan. 2021.

[3] D. R. Prapti, A. R. M. Shariff, H. C. Man, N. M. Ramli, T. Perumal, and M. Shariff, "Internet of Things (IoT)-based aquaculture: An overview of IoT application on water quality monitoring," *Rev. Aquaculture*, vol. 14, no. 2, pp. 979–992, Mar. 2022.

[4] M. Manoj, V. D. Kumar, M. Arif, E.-R. Bulai, P. Bulai, and O. Geman, "State of the art techniques for water quality monitoring systems for fish ponds using IoT and underwater sensors: A review," *Sensors*, vol. 22, no. 6, p. 2088, Mar. 2022.

[5] F. Alam, A. Gupta, A. Saha, and S. Md. Salimullah, "Establishing an Internet of Things (IoT)-enabled solar-powered smart water treatment system," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2023, pp. 1–6.

[6] F. K. Shaikh, M. A. Memon, N. A. Mahoto, S. Zeadally, and J. Nebhen, "Artificial intelligence best practices in smart agriculture," *IEEE Micro*, vol. 42, no. 1, pp. 17–24, Jan. 2022, doi: 10.1109/MM.2021.3121279.

[7] F. K. Shaikh, S. Karim, S. Zeadally, and J. Nebhen, "Recent trends in Internet-of-Things-Enabled sensor technologies for smart agriculture," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23583–23598, Dec. 2022, doi: 10.1109/JIOT.2022.3210154.

[8] T. Karanisa, Y. Achour, A. Ouammi, and S. Sayadi, "Smart greenhouses as the path towards precision agriculture in the food-energy and water nexus: Case study of Qatar," *Environ. Syst. Decisions*, vol. 42, no. 4, pp. 521–546, 2022.

[9] P. K. Kashyap, S. Kumar, A. Jaiswal, M. Prasad, and A. H. Gandomi, "Towards precision agriculture: IoT-enabled intelligent irrigation systems using deep learning neural network," *IEEE Sensors J.*, vol. 21, no. 16, pp. 17479–17491, Aug. 2021.

[10] H. Mehmood, D. Liao, and K. Mahadeo, "A review of artificial intelligence applications to achieve water-related sustainable development goals," in *Proc. IEEE/ITU Int. Conf. Artif. Intell. Good (AI4G)*, Sep. 2020, pp. 135–141, doi: 10.1109/AI4G50087.2020.9311018.

[11] C. V. Chinnappan, A. D. J. William, S. K. C. Nidamanuri, S. Jayalakshmi, R. Bogani, P. Thanapal, S. Syed, B. Venkateswarlu, and J. A. I. S. Masood, "IoT-enabled chlorine level assessment and prediction in water monitoring system using machine learning," *Electronics*, vol. 12, no. 6, p. 1458, Mar. 2023.

[12] N. Islam and K. Irshad, "Artificial ecosystem optimization with deep learning enabled water quality prediction and classification model," *Chemosphere*, vol. 309, Dec. 2022, Art. no. 136615.

[13] T. Tamilselvi, R. Saravanakumar, R. Arunkumar, and K. Revathi, "H2O caliber: An IoT enabled surface water pollutant assessment system with deep learning," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, no. 2, pp. 531–536, 2023.

[14] S. Kimothi, A. Thapliyal, S. V. Akram, R. Singh, A. Gehlot, H. G. Mohamed, D. Anand, M. Ibrahim, and I. D. Noya, "Big data analysis framework for water quality indicators with assimilation of IoT and ML," *Electronics*, vol. 11, no. 13, p. 1927, Jun. 2022.

[15] F. Akhter, H. R. Siddiquei, M. E. E. Alahi, K. P. Jayasundera, and S. C. Mukhopadhyay, "An IoT-enabled portable water quality monitoring system with MWCNT/PDMS multifunctional sensor for agricultural applications," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 14307–14316, Aug. 2022, doi: 10.1109/JIOT.2021.3069894.

[16] J. O. Ighalo, A. G. Adeniyi, and G. Marques, "Internet of Things for water quality monitoring and assessment: A comprehensive review," in *Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications*. Cham, Switzerland: Springer, 2021, pp. 245–259.

[17] F. K. Shaikh and S. Zeadally, "Energy harvesting in wireless sensor networks: A comprehensive review," *Renew. Sustain. Energy Rev.*, vol. 55, pp. 1041–1054, Mar. 2016.

[18] S. Zeadally, F. K. Shaikh, A. Talpur, and Q. Z. Sheng, "Design architectures for energy harvesting in the Internet of Things," *Renew. Sustain. Energy Rev.*, vol. 128, Aug. 2020, Art. no. 109901.

[19] J.-K. Kang, D. Lee, K. E. Muambo, J.-W. Choi, and J.-E. Oh, "Development of an embedded molecular structure-based model for prediction of micropollutant treatability in a drinking water treatment plant by machine learning from three years monitoring data," *Water Res.*, vol. 239, Jul. 2023, Art. no. 120037.

[20] R. Huang, C. Ma, J. Ma, X. Huangfu, and Q. He, "Machine learning in natural and engineered water systems," *Water Res.*, vol. 205, Oct. 2021, Art. no. 117666.

[21] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches," *Water Res.*, vol. 229, Feb. 2023, Art. no. 119422.

[22] F. Muharemi, D. Logofătu, and F. Leon, "Machine learning approaches for anomaly detection of water quality on a real-world data set," *J. Inf. Telecommun.*, vol. 3, no. 3, pp. 294–307, 2019.

[23] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "Performance analysis of the water quality index model for predicting water state using machine learning techniques," *Process Saf. Environ. Protection*, vol. 169, pp. 808–828, Jan. 2023.

[24] Ma. Pedro-Monzonís, A. Solera, J. Ferrer, T. Estrela, and J. Paredes-Arquiola, "A review of water scarcity and drought indexes in water resources planning and management," *J. Hydrol.*, vol. 527, pp. 482–493, Aug. 2015.

[25] B. Sarker, K. N. Keya, F. I. Mahir, K. M. Nahiun, S. Shahida, and R. A. Khan, "Surface and ground water pollution: Causes and effects of urbanization and industrialization in South Asia," *Sci. Rev.*, vol. 7, no. 73, pp. 32–41, Jul. 2021.

[26] R. P. Schwarzenbach, T. Egli, T. B. Hofstetter, U. V. Gunten, and B. Wehrli, "Global water pollution and human health," *Annu. Rev. Environ. Resour.*, vol. 35, pp. 109–136, Nov. 2010.

[27] M. Kumar, T. Singh, M. K. Maurya, A. Shivhare, A. Raut, and P. K. Singh, "Quality assessment and monitoring of river water using IoT infrastructure," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10280–10290, Jun. 2023.

[28] S. Tian, H. Guo, W. Xu, X. Zhu, B. Wang, Q. Zeng, Y. Mai, and J. J. Huang, "Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms," *Environ. Sci. Pollut. Res.*, vol. 30, no. 7, pp. 18617–18630, Oct. 2022.

[29] A. A. Nasser, M. Z. Rashad, and S. E. Hussein, "A two-layer water demand prediction system in urban areas based on micro-services and LSTM neural networks," *IEEE Access*, vol. 8, pp. 147647–147661, 2020.

[30] J. Morón-López, M. C. Rodríguez-Sánchez, F. Carreño, J. Vaquero, Á. G. Pompa-Pernía, M. Mateos-Fernández, and J. A. P. Aguilar, "Implementation of smart buoys and satellite-based systems for the remote monitoring of harmful algae Bloom in inland waters," *IEEE Sensors J.*, vol. 21, no. 5, pp. 6990–6997, Mar. 2020.

[31] E. Menasalvas, N. Swoboda, A. Moreno, A. Metzger, A. Rothweiler, N. Pavlopoulou, and E. Curry, "Recognition of formal and non-formal training in data science," in *The Elements Big Data Value*. Cham, Switzerland: Springer, 2021, p. 311.

[32] W. Dong and Q. Yang, "Data-driven solution for optimal pumping units scheduling of smart water conservancy," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1919–1926, Mar. 2020, doi: 10.1109/JIOT.2019.2963250.

[33] X. Wang, X. Gao, Y. Zhang, X. Fei, Z. Chen, J. Wang, Y. Zhang, X. Lu, and H. Zhao, "Land-cover classification of coastal wetlands using the RF algorithm for Worldview-2 and Landsat 8 images," *Remote Sens.*, vol. 11, no. 16, p. 1927, Aug. 2019.

[34] H. Ghorbani, D. A. Wood, A. Choubineh, A. Tatar, P. G. Abarghoyi, M. Madani, and N. Mohamadian, "Prediction of oil flow rate through an orifice flow meter: Artificial intelligence alternatives compared," *Petroleum*, vol. 6, no. 4, pp. 404–414, Dec. 2020.

[35] K. DeMedeiros, A. Hendawi, and M. Alvarez, "A survey of AI-based anomaly detection in IoT and sensor networks," *Sensors*, vol. 23, no. 3, p. 1352, Jan. 2023.

[36] M. Miller, A. Kisiel, D. Cembrowska-Lech, I. Durlik, and T. Miller, "IoT in water quality monitoring-are we really here?" *Sensors*, vol. 23, no. 2, p. 960, 2023.

[37] R. Haggerty, J. Sun, H. Yu, and Y. Li, "Application of machine learning in groundwater quality modeling—A comprehensive review," *Water Res.*, vol. 233, Apr. 2023, Art. no. 119745.

[38] H. S. Barjouei, H. Ghorbani, N. Mohamadian, D. A. Wood, S. Davoodi, J. Moghadasi, and H. Saberi, "Prediction performance advantages of deep machine learning algorithms for two-phase flow rates through wellhead chokes," *J. Petroleum Explor. Prod. Technol.*, vol. 11, no. 3, pp. 1233–1261, Mar. 2021.

[39] A. Bhardwaj, V. Dagar, M. O. Khan, A. Aggarwal, R. Alvarado, M. Kumar, M. Irfan, and R. Proshad, "Smart IoT and machine learning-based framework for water quality assessment and device component monitoring," *Environ. Sci. Pollut. Res.*, vol. 29, no. 30, pp. 46018–46036, Jun. 2022.

[40] S. Balaji, K. Nathani, and R. Santhakumar, "IoT technology, applications and challenges: A contemporary survey," *Wireless Pers. Commun.*, vol. 108, no. 1, pp. 363–388, Sep. 2019.

[41] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for Internet of Things (IoT) security," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1646–1685, 3rd Quart., 2020.

[42] R. Boutaba, M. A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. M. Caicedo, "A comprehensive survey on machine learning for networking: Evolution, applications and research opportunities," *J. Internet Services Appl.*, vol. 9, no. 1, pp. 1–99, Dec. 2018.

[43] R. A. Koestoer, Y. A. Saleh, I. Roihan, and Harinaldi, "A simple method for calibration of temperature sensor DS18B20 waterproof in oil bath based on Arduino data acquisition system," in *Proc. AIP Conf.*, 2019, Art. no. 020006.

[44] T. H. Nasution, S. Dika, E. P. Sinulingga, K. Tanjung, and L. A. Harahap, "Analysis of the use of SEN0161 pH sensor for water in goldfish ponds," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 851, no. 1, May 2020, Art. no. 012053.

[45] J. Tomperi, A. Isokangas, T. Tuuttila, and M. Paavola, "Functionality of turbidity measurement under changing water quality and environmental conditions," *Environ. Technol.*, vol. 43, no. 7, pp. 1093–1101, Mar. 2022.

[46] A. A. A. Maliki, A. Chabuk, M. A. Sultan, B. M. Hashim, H. M. Hussain, and N. Al-Ansari, "Estimation of total dissolved solids in water bodies by spectral indices case study: Shatt Al-Arab river," *Water, Air, Soil Pollut.*, vol. 231, no. 9, pp. 1–11, Sep. 2020.

[47] D. T. Vo, X. P. Nguyen, T. D. Nguyen, R. Hidayat, T. T. Huynh, and D. T. Nguyen, "A review on the Internet of thing (IoT) technologies in controlling ocean environment," *Energy Sources, A, Recovery, Utilization, Environ. Effects*, pp. 1–19, Jul. 2021.

[48] R. Bhateria and D. Jain, "Water quality assessment of lake water: A review," *Sustain. Water Resour. Manage.*, vol. 2, no. 2, pp. 161–173, Jun. 2016.

[49] L. Chen, Y. Xu, F. Xu, Q. Hu, and Z. Tang, "Balancing the trade-off between cost and reliability for wireless sensor networks: A multi-objective optimized deployment method," *Int. J. Speech Technol.*, vol. 53, no. 8, pp. 9148–9173, Apr. 2023.

[50] M. Hamzei, S. Khandagh, and N. Jafari Navimipour, "A quality-of-service-aware service composition method in the Internet of Things using a multi-objective fuzzy-based hybrid algorithm," *Sensors*, vol. 23, no. 16, p. 7233, Aug. 2023.

[51] *Store Arduino Arduino*, Arduino, Arduino LLC, Somerville, MA, USA, 2015, p. 372.

[52] A. Augustin, J. Yi, T. Clausen, and W. Townsley, "A study of LoRa: Long range & low power networks for the Internet of Things," *Sensors*, vol. 16, no. 9, p. 1466, Sep. 2016.

[53] L. Vangelista, A. Zanella, and M. Zorzi, "Long-range IoT technologies: The dawn of LoRa," in *Future Access Enablers of Ubiquitous and Intelligent Infrastructures*. Cham, Switzerland: Springer, 2015, pp. 51–58.

[54] *Developing Drinking-Water Quality Regulations Standards*, World Health Organization, Geneva, Switzerland, 2018.

[55] M. Arshad and A. Shakoor, "Irrigation water quality," *Water Int*, vol. 12, nos. 1–2, pp. 145–160, 2017.

[56] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[57] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment," *Water Res.*, vol. 219, Jul. 2022, Art. no. 118532.

[58] S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, "Water quality assessment in terms of water quality index," *Amer. J. Water Resour.*, vol. 1, no. 3, pp. 34–38, Oct. 2020.

[59] G. Srivastava and P. Kumar, "Water quality index with missing parameters," *Int. J. Res. Eng. Technol.*, vol. 5, no. 4, pp. 609–614, Apr. 2013.

[60] H. Aydin, F. Ustaoğlu, Y. Tepe, and E. N. Soylu, "Assessment of water quality of streams in northeast Turkey by water quality index and multiple statistical methods," *Environ. Forensics*, vol. 22, nos. 1–2, pp. 270–287, 2021.

[61] A. Rahman, *Statistics for Data Science and Policy Analysis*. Cham, Switzerland: Springer, 2020.

[62] A. Rahman, "Statistics-based data preprocessing methods and machine learning algorithms for big data analysis," *Int. J. Artif. Intell.*, vol. 17, no. 2, pp. 44–65, 2019.

[63] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagation for classification," *Int. J. Comput. Theory Eng.*, vol. 3, no. 1, p. 1793, 2011.

[64] E. A. Curtis, C. Comiskey, and O. Dempsey, "Importance and use of correlational research," *Nurse Researcher*, vol. 23, no. 6, pp. 20–25, Jul. 2016.

[65] A. Ganti, "Correlation coefficient," *Corp. Financ. Account*, vol. 9, pp. 145–152, May 2020.

[66] S. A. Osmani, B. K. Banik, and H. Ali, "Integrating fuzzy logic with Pearson correlation to optimize contaminant detection in water distribution system with uncertainty analyses," *Environ. Monitor. Assessment*, vol. 191, no. 7, pp. 1–15, Jul. 2019.

[67] D. V. V. Prasad, L. Y. Venkataramana, P. S. Kumar, G. Prasannamedha, S. Harshana, S. J. Srividya, K. Harrinei, and S. Indraganti, "Analysis and prediction of water quality using deep learning and auto deep learning techniques," *Sci. Total Environ.*, vol. 821, May 2022, Art. no. 153311.

[68] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: What does it estimate and how well does it do it?" *J. Amer. Stat. Assoc.*, pp. 1–12, 2023, doi: 10.1080/01621459.2023.2197686.

[69] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," *Ecolog. Monographs*, vol. 93, no. 1, Feb. 2023, Art. no. e1557.

[70] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D: Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306.

[71] F. Zhang and L. J. O'Donnell, "Support vector regression," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 123–140.

[72] F. Günther and S. Fritsch, "NeuralNet: Training of neural networks," *R J.*, vol. 2, no. 1, p. 30, 2010.

[73] Z. Yang and E. E. Mehmed, "Artificial neural networks in freight rate forecasting," *Maritime Econ. Logistics*, vol. 21, no. 3, pp. 390–414, Sep. 2019.

[74] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.
[75] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, and H. Cho, "Xgboost: Extreme gradient boosting," *R Package Version*, vol. 1, no. 4, pp. 1–4, Aug. 2015.
[76] W. Alajali, W. Zhou, S. Wen, and Y. Wang, "Intersection traffic prediction using decision tree models," *Symmetry*, vol. 10, no. 9, p. 386, Sep. 2018.
[77] J. R. Quinlan, "Decision trees and decision-making," *IEEE Trans. Syst. Man, Cybern.*, vol. 20, no. 2, pp. 339–346, Mar. 1990.
[78] A. Liaw et al., "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[79] M. G. Uddin, S. Nash, M. T. M. Diganta, A. Rahman, and A. I. Olbert, "Robust machine learning algorithms for predicting coastal water quality index," *J. Environ. Manage.*, vol. 321, Nov. 2022, Art. no. 115923.
[80] Z. Xiong, Y. Cui, Z. Liu, Y. Zhao, M. Hu, and J. Hu, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation," *Comput. Mater. Sci.*, vol. 171, Jan. 2020, Art. no. 109203.
[81] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Proc. Australas. Joint Conf. Artif. Intell.* Hobart, TS, Australia: Springer, Dec. 2006, pp. 1015–1021, 2006.
[82] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-Score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Santiago de Compostela, Spain: Springer, Mar. 2005, pp. 345–359.
[83] M. Najafzadeh and S. Basirian, "Evaluation of river water quality index using remote sensing and artificial intelligence models," *Remote Sens.*, vol. 15, no. 9, p. 2359, Apr. 2023.
[84] B.-B. Jia, J.-Y. Liu, J.-Y. Hang, and M.-L. Zhang, "Learning label-specific features for decomposition-based multi-class classification," *Frontiers Comput. Sci.*, vol. 17, no. 6, pp. 1–10, Dec. 2023.
[85] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes," *Pattern Recognit.*, vol. 44, no. 8, pp. 1761–1776, Aug. 2011.
[86] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang, "Water quality prediction based on machine learning and comprehensive weighting methods," *Entropy*, vol. 25, no. 8, p. 1186, Aug. 2023.
[87] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, no. 11, p. 2210, 2019.
[88] T. H. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water quality prediction using artificial intelligence algorithms," *Appl. Bionics Biomech.*, vol. 2020, pp. 1–12, Dec. 2020.
[89] R. Gai and H. Zhang, "Prediction model of agricultural water quality based on optimized logistic regression algorithm," *EURASIP J. Adv. Signal Process.*, vol. 2023, no. 1, p. 21, Feb. 2023.
[90] Z. Ahmad, N. A. Rahim, A. Bahadori, and J. Zhang, "Improving water quality index prediction in perak river basin Malaysia through a combination of multiple neural networks," *Int. J. River Basin Manage.*, vol. 15, no. 1, pp. 79–87, Jan. 2017.
[91] M. Sakizadeh, "Artificial intelligence for the prediction of water quality index in groundwater systems," *Model. Earth Syst. Environ.*, vol. 2, no. 1, pp. 1–9, Mar. 2016.
[92] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," *Mar. Pollut. Bull.*, vol. 64, no. 11, pp. 2409–2420, Nov. 2012.
[93] V. Ranković, J. Radulović, I. Radojević, A. Ostojić, and L. Čomić, "Neural network modeling of dissolved oxygen in the Gruža Reservoir, Serbia," *Ecolog. Model.*, vol. 221, no. 8, pp. 1239–1244, Apr. 2010.
[94] J. I. Ubah, L. C. Orakwe, K. N. Ogbu, J. I. Awu, I. E. Ahaneku, and E. C. Chukwuma, "Forecasting water quality parameters using artificial neural network for irrigation purposes," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, Art. no. 24438.
[95] M. S. Mazorchuk, T. S. Vakulenko, A. O. Bychko, O. H. Kuzminska, and O. V. Prokhorov, "Cloud technologies and learning analytics: Web application for Pisa results analysis and visualization," in *Proc. CTE Workshop*, vol. 8, Mar. 2021, pp. 484–494.
[96] A. Protopsaltis, P. Sarigiannidis, D. Margounakis, and A. Lytos, "Data visualization in Internet of Things: Tools, methodologies, and challenges," in *Proc. 15th Int. Conf. Availability, Rel. Secur.* New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–11.
[97] K. Aurangzeb, F. Akmal, M. A. Khan, M. Sharif, and M. Y. Javed, "Advanced machine learning algorithm based system for crops leaf diseases recognition," in *Proc. 6th Conf. Data Sci. Mach. Learn. Appl. (CDMA)*, Mar. 2020, pp. 146–151.
[98] M. Alhussein, K. Aurangzeb, and S. I. Haider, "Hybrid CNN-LSTM model for short-term individual household load forecasting," *IEEE Access*, vol. 8, pp. 180544–180557, 2020.
[99] A. Khan, S. Aslam, K. Aurangzeb, M. Alhussein, and N. Javaid, "Multiscale modeling in smart cities: A survey on applications, current trends, and challenges," *Sustain. Cities Soc.*, vol. 78, Mar. 2022, Art. no. 103517.

MUSHTAQUE AHMED RAHU received the B.Eng. degree in electronic engineering and the M.Eng. degree in industrial automation and control from the Quaid-e-Awam University of Engineering, Science and Technology (QUEST), Nawabshah, Pakistan, in April 2010 and 2016, respectively, where he is currently pursuing the Ph.D. degree. His research interests include the Internet of Things, wireless sensor networks, machine learning, and smart agriculture. He is a Lifetime Member of the Pakistan Engineering Council.

ABDUL FATTAH CHANDIO received the B.Eng. degree in electronic engineering from the Mehran University of Engineering and Technology, Jamshoro, the M.Eng. degree in telecom engineering from the National University of Science and Technology (Military College of Signals Campus), and the Ph.D. degree in information sciences from the Beijing Institute of Technology, Beijing, China. He is working as Dean Faculty of Electrical, Electronics and Computer Systems Engineering, Quaid-e-Awam University of Engineering, Science and Technology (QUEST), Nawabshah, Pakistan. His research interests include spatial data mining (hazard mitigation, environmental protection, and resource distribution), the use of optical signal processing in computer hardware (holography and optical interfacing), and solutions of computer communications and internet.

KHURSHEED AURANGZEB (Senior Member, IEEE) received the B.S. degree in computer engineering from the COMSATS Institute of Information Technology Abbottabad, Pakistan, in 2006, the M.S. degree in electrical engineering (system on chip design) from Linköping University, Sweden, in 2009, and the Ph.D. degree in electronics design from Mid Sweden University, Sweden, in June 2013. He is currently an Associate Professor with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He has obtained more than 15 years of excellent experience as an instructor and a researcher in data analytics, machine/deep learning, signal processing, electronics circuits/systems, and embedded systems. He has been involved in many research projects as a principal investigator and a co-principal investigator. He has authored or coauthored more than 90 publications, including IEEE/ACM/Springer/Hindawi/MDPI journals, and flagship conference papers. His research interests include embedded systems, computer architecture, signal processing, wireless sensor networks, communication, and camera-based sensor networks, with an emphasis on big data and machine/deep learning with applications in smart grids, precision agriculture, and healthcare.

SARANG KARIM received the B.Eng. degree in electronic engineering from the Mehran University of Engineering and Technology (MUET), Jamshoro, Pakistan, in April 2011, and the M.Eng. degree in electronic systems engineering from the Institute of Information and Communication Technologies (IICT), MUET, in 2015. He is currently pursuing the Ph.D. degree with the IICT, MUET. He was attached with ETSI, Universidad de Málaga, Málaga, Spain, as a Mobility Researcher, from September 2017 to February 2018. He is currently a Lecturer with the Department of Telecommunication Engineering, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan. He published more than 15 research papers in reputed journals and conference proceedings. His research interests include the Internet of Things, wireless sensor networks, underwater wireless sensor networks, and smart agriculture. He is a Lifetime Member of the Pakistan Engineering Council.

MUSAED ALHUSSEIN received the B.S. degree in computer engineering from King Saud University (KSU), Riyadh, Saudi Arabia, in 1988, and the M.S. and Ph.D. degrees in computer science and engineering from the University of South Florida, Tampa, FL, USA, in 1992 and 1997, respectively. Since 1997, he has been a Faculty Member with the Computer Engineering Department, College of Computer and Information Science, KSU. He is currently a Professor with the Department of Computer Engineering, College of Computer and Information Sciences, KSU. He is also the Founder and the Director of the Embedded Computing and Signal Processing Research (ECASP) Laboratory. Recently, he has been successful in winning a research project in the area of AI for healthcare, which is funded by the Ministry of Education at Saudi Arabia. His research interests include typical computer architecture and signal processing topics with an emphasis on big data, machine/deep learning, VLSI testing and verification, embedded and pervasive computing, cyber-physical systems, mobile cloud computing, big data, eHealthcare, and body area networks.

MUHAMMAD SHAHID ANWAR received the M.Sc. degree in telecommunications technology from Aston University, Birmingham, U.K., in 2012, and the Ph.D. degree in information and communication engineering from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2021. He is currently an Assistant Professor with the Department of AI and Software, Gachon University, Seongnam, South Korea. He has published more than 30 research papers in reputed international journals and conferences. His research interests include 360-degree videos, virtual reality (VR), AR, metaverse, and quality of experience (QoE) evaluations of immersive media. He is focusing on deep learning-based VR video evaluations and developed several machine learning-based QoE prediction models.