International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Machine Learning Adopted Potability Assessment for Safe Drinking Water

S. Abhishek Yadav[1], Borada Yugender[2], Telaboina Sushma[2], Kanchanapally Vighnan[2], Domala Manoj Kumar[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering (AI & ML)
[1,2]Malla Reddy Engineering College and Management Science, Kistapur, Medchal-50140l, Hyderabad, Telangana, India

ABSTRACT

Access to clean and safe drinking water is a fundamental human right. Over the years, various methods and standards have been established to ensure the potability (safety for consumption) of water. Traditional methods involve extensive laboratory testing, which can be time-consuming, expensive, and may not always be practical, especially in remote or resource-limited areas. Machine learning offers a promising approach to automate and enhance the process of water potability assessment. Machine learning-based systems can provide a faster and more cost-effective means of evaluating water potability, enabling timely interventions to ensure the health and well-being of communities. The project, "Machine Learning-Based Potability Assessment of Drinking Water for Human Consumption," seeks to leverage the power of advanced machine learning algorithms to enhance the efficiency and accessibility of water quality assessment. By training models on comprehensive datasets of water quality parameters and corresponding potability labels, this research aims to develop a system capable of autonomously and accurately classifying drinking water. The integration of machine learning techniques allows for the extraction of complex relationships and patterns within the data, enabling precise potability assessments. This advancement holds great promise for ensuring the availability of safe and clean drinking water for communities worldwide, especially in regions facing water quality challenges.

Keywords: Water quality monitoring, Data analytics, Predictive analytics, Machine learning, Potability assessment.

1. Introduction

Clean and safe drinking water is an indispensable cornerstone of a healthy society, and ensuring its accessibility is a fundamental human right. Throughout history, mankind has grappled with the challenge of guaranteeing the potability of water, employing diverse methods and standards. Traditional approaches to assessing water safety have primarily relied on meticulous laboratory testing, a process characterized by its accuracy but plagued by issues of time consumption, financial burden, and practicality, particularly in remote or resource-scarce regions. As the global community confronts escalating concerns regarding water quality, driven by factors such as pollution, burgeoning populations, and the pervasive impact of climate change, the imperative for efficient and dependable methods to evaluate the safety of drinking water becomes increasingly apparent. The conventional methodologies employed in water potability assessment are deeply rooted in extensive chemical analyses carried out within the confines of laboratories. While undeniably precise, these methods suffer from inherent drawbacks that hinder their universal applicability. The time-intensive nature of laboratory testing, coupled with its resource-intensive demands, poses significant challenges to its feasibility, particularly in remote or underserved areas where accessibility to advanced infrastructure is limited. The pressing challenge at hand is to transcend the limitations of traditional assessment methods by harnessing the power of machine learning, an innovative approach that holds the potential to revolutionize and streamline the evaluation of water potability.

The paradigm shifts towards integrating machine learning into the assessment of drinking water potability is motivated by the need for a faster, more cost-effective, and widely applicable means of ensuring water safety. The project, titled "Machine Learning-Based Potability Assessment of Drinking Water for Human Consumption," emerges as a beacon of hope in this pursuit. At its core, the project seeks to leverage advanced machine learning algorithms to enhance the efficiency and accessibility of water quality assessment, ultimately contributing to the global goal of providing safe and clean drinking water to all communities. The crux of the initiative lies in the development of a machine learning model adept at accurately assessing the potability of drinking water based on a myriad of chemical and physical parameters. This intricate process involves the meticulous processing and analysis of extensive water quality data to ascertain whether it adheres to the requisite standards for human consumption. The significance of this endeavour cannot be overstated, especially against the backdrop of increasing environmental challenges that threaten the availability of safe drinking water on a global scale.

The motivation behind adopting machine learning as the cornerstone of this assessment lies in its ability to transcend the constraints of traditional methods. By training models on comprehensive datasets encompassing a spectrum of water quality parameters and their corresponding potability labels, the research endeavors to develop a system capable of autonomously and accurately classifying drinking water. Unlike traditional methods, machine learning techniques allow for the extraction of intricate relationships and patterns within the data, enabling nuanced and precise assessments of water potability The core challenge addressed by the project, "Machine Learning-Based Potability Assessment of Drinking Water for Human Consumption," lies in the inherent limitations and inefficiencies of traditional methods employed in water potability assessment. The reliance on laborious and time-consuming laboratory testing, while undeniably accurate, poses significant practical hurdles, particularly in remote or resource-scarce areas. The need to adapt to evolving global challenges related to water quality, spurred by factors such as pollution, population growth, and climate change, necessitates a departure from conventional approaches. The problem at hand is the imperative to develop a more expedient, cost-effective, and universally applicable method for evaluating the safety of drinking water.

The rigidities of existing assessment methods, coupled with the escalating threats to water resources, underscore the urgency of adopting a transformative approach. The machine learning paradigm, in this context, becomes a solution to overcome the limitations of traditional methods. The primary problem is to harness the power of machine learning algorithms to autonomously and accurately classify drinking water based on diverse chemical and physical parameters. The intricate nature of water quality parameters, influenced by dynamic real-world conditions, adds complexity to the problem. Adapting machine learning models to comprehend and adapt to this dynamic nature represents a key challenge. Moreover, ethical considerations surrounding the deployment of machine learning in critical public health areas necessitate careful attention. The problem statement encapsulates the need for a systematic departure from established norms, driven by the urgency to provide efficient and reliable methods for safeguarding the fundamental human right to access safe and clean drinking water, especially in the face of pressing environmental challenges.

2. Literature Survey

Water is a fundamental natural resource for all life forms on planet Earth. Safe water should be free from harmful chemical substances or microorganisms at concentrations that cause health problems, according to the recommendations of the World Health Organization (WHO) [1]. Rivers and lakes are considered the main sources of freshwater and represent one of the most important water resources for various uses, such as drinking, agriculture, industry, and domestic needs. They resemble lifelines for communities and play a crucial role in social, economic, and environmental development [2].

However, these water bodies are severely depleted due to excessive human activities, such as manufacturing, urbanization, and population growth. Surface water sources including rivers and lakes have been subjected to widespread pollution from various sources, according to the United Nations Environment Programme [3]. In addition, poor water resource management and climate change have caused a decline in water quality in recent decades, leading to surface water pollution [4]. The surface water quality in a region largely depends on the nature and level of various human activities in the relevant watersheds.

The chemical, physical, and biological compositions of surface water are subject to numerous effects, including natural effects such as rainfall, watershed geography, atmosphere, and geology, as well as human effects such as industrial, agricultural, and household activities [5]. Increasing surface water pollution leads to the deterioration in water quality, threatens human health, affects the balance of the aquatic ecosystem, and hinders economic development and social progress [6]. According to a report by the WHO, polluted water causes about 80% of human diseases. When groundwater is polluted, its quality can be restored by stopping the flow of pollutants from the source [7]. Therefore, it is essential to continuously monitor the quality of surface and groundwater and improve the methods and means to protect them.

The water quality index (WQI) is used to assess and summarize the overall water quality of water [7]. It takes into account various physical, chemical, and biological parameters, including temperature, pH, dissolved oxygen, turbidity, and levels of pollutants such as nutrients and contaminants. The WQI provides a numerical value or rating that helps determine the health and suitability of water for different uses, such as drinking, recreation, or aquatic life. Higher WQI values generally indicate better water quality, while lower values suggest poorer water conditions. This index helps decision makers to take effective measures to manage water resources and maintain their quality [6,8]. The formulation and use of quality indices have been strongly supported by organizations responsible for water supply and pollution control. Nevertheless, the utilization of the WQI to evaluate groundwater and surface water quality was limited for a long time due to the lack of sufficient data and appropriate statistical and modeling methods.

In recent years, machine learning (ML) techniques have been widely used to evaluate water quality, including estimating the WQI [9]. These techniques have proven powerful tools for modeling complex linear and nonlinear relationships in environmental and water resource research [10]. The application of multivariate statistical methods, such as multiple linear regression (MLR), cluster analysis (CA), principal components analysis (PCA), factor analysis (FA), and discriminant analysis (DA), is useful in reducing the complexity of large water quality data sets (reducing the number of variables) without losing the original information [11]. Applying these statistical techniques helps interpret complex data to better understand the environmental water quality status and identify potential sources or factors that affect water systems, in addition to providing a quick solution to pollution problems for simple and cost-effective water quality assessment.

A literature review shows that each ML algorithm has its strengths and weaknesses, and its behavior depends on the water quality input variables in different study areas [12, 13]. Gupta and Gupta investigated the health status of the Damodar River in India for drinking purposes using the WQI method. They analyzed eleven water quality parameters from ten monitoring sites along the river and applied an MLR model to predict WQI. The results showed that river health varied between good and unfit categories. In addition, it identified biochemical oxygen demand (BOD), total coliform (TC), and iron (Fe) as the primary factors affecting WQI values, and the MLR model was found to be effective for evaluating river health for efficient river management. The model exhibited a strong fit, indicating

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

a robust relationship between the identified factors and the WQI values. The results underscore the potential of the MLR model as a valuable tool for evaluating river health [14].

3. Proposed Methodology

3.1 Overview

Throughout this comprehensive research procedure, the seamless integration of Flask server, HTML-based UI, user-provided data, and machine learning models underscores a holistic approach to water potability classification. The fusion of technological innovation and user-centric design ensures not only the technical accuracy of the system but also its accessibility and relevance in real-world contexts. The binary classification models, KNN and DTC-Adaboost, represent a strategic selection to capitalize on diverse machine learning techniques, enhancing the overall robustness of the classification system. The performance estimation phase acts as a pivotal checkpoint, guiding the refinement process and steering the project towards optimal functionality. The culmination in data prediction from test data serves as the ultimate validation, affirming the system's efficacy in contributing to the crucial domain of water quality assessment. In essence, this research work embodies a synergistic fusion of technological prowess and user engagement, offering a promising solution to the pressing challenge of ensuring the potability of drinking water through innovative machine learning-based classification. Figure 1 shows the proposed system model.

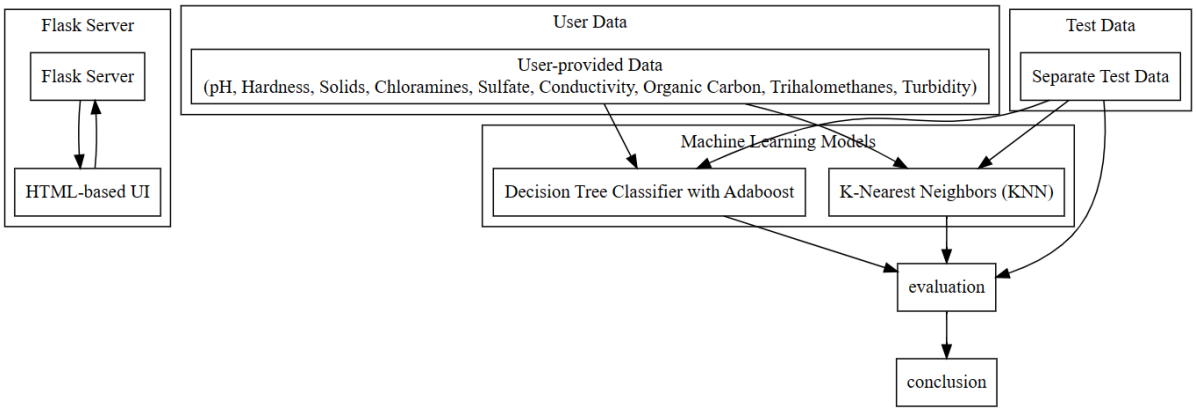

Figure 1: Proposed system model.

The comprehensive research work on water potability classification unfolds through a systematic and multifaceted procedure, integrating both technological and user-interface elements. The endeavor begins with the creation of a Flask server, establishing the foundational infrastructure for hosting the subsequent components of the project. This server acts as the backbone, facilitating communication and coordination among various modules.

The second phase of the research involves the creation of an HTML-based user interface (UI), a crucial element in ensuring user-friendly interaction with the water potability classification system. The UI serves as the portal through which users input relevant data for assessment. This interface is designed to be intuitive and accessible, catering to diverse user backgrounds and ensuring a seamless experience in contributing data for water quality evaluation.

With the UI in place, the third step involves applying input data to the system. The data, comprising key parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, is collected from users via the interface. This step is pivotal as it constitutes the raw material upon which the machine learning models will be trained and tested. The

incorporation of user-provided data adds a dynamic and real-world dimension to the system, enhancing its applicability across diverse scenarios.

Moving forward, the fourth and fifth steps focus on the training of machine learning models. The research incorporates two distinct models for binary classification—K-Nearest Neighbors (KNN) and a Decision Tree Classifier (DTC) boosted with Adaboost. The KNN model leverages proximity-based classification, while the DTC-Adaboost classifier combines the strengths of decision trees and boosting algorithms to enhance predictive accuracy. The training process involves exposing these models to a comprehensive dataset, encompassing a range of water quality parameters and corresponding potability labels. This phase is crucial for enabling the models to discern intricate patterns and relationships within the data, establishing a foundation for accurate classification.

Post-training, the sixth step revolves around performance estimation. This critical evaluation assesses the efficacy of the trained models in accurately classifying water potability. Metrics such as accuracy, precision, recall, and F1 score are employed to gauge the models' effectiveness in binary classification. This step serves as a litmus test for the robustness and reliability of the machine learning models, informing subsequent refinements and improvements.

The final step, data prediction from test data, involves subjecting the trained models to real-world scenarios. The models are tested with a separate set of data not used during the training phase to simulate real-world conditions. This phase is instrumental in validating the models' ability to generalize and make accurate predictions beyond the training dataset. It also provides insights into the models' adaptability to variations in water quality parameters, reinforcing the practical utility of the classification system.

3.2 Flask with HTML

The HTTP protocol, or Hypertext Transfer Protocol, serves as the foundation for communication on the World Wide Web. It allows for the exchange of information between a client (such as a web browser) and a server. When a user enters a website's address into the browser's address bar and presses enter, an HTTP request is generated and sent to the server associated with that address. The server then interprets the request, processes it, and sends back an HTTP response containing the requested information, which is then displayed on the user's browser. In the context of web development, particularly with the Python programming language, Flask is a micro-framework that simplifies the process of creating web applications. Flask facilitates the handling of incoming HTTP requests, allowing developers to focus on defining how the application responds to those requests. In essence, Flask acts as the bridge between the user's browser and the server, managing the flow of information and ensuring that the appropriate responses are generated.

The Flask framework operates by defining routes, which specify the URL paths that the application will respond to. In the provided code example, a basic Flask application is created. The **@app.route("/")** decorator indicates that when a user accesses the root URL (e.g., "http://localhost:5000/"), the associated function (**home()** in this case) should be executed. The function returns the string "Hello, World!" as a response, which is then displayed on the user's browser. Furthermore, Flask enables the integration of HTML templates to enhance the appearance of web pages. HTML files are stored in a folder named "templates," and the **render_template** function is employed to render these templates. The provided code introduces two additional routes, "/salvador" and "/about," each associated with a specific HTML template ("home.html" and "about.html," respectively). These templates are rendered when the corresponding routes are accessed, creating a more dynamic and visually appealing user experience.

To illustrate, the "home.html" template includes basic HTML structure with a heading and a paragraph. The "about.html" template provides information about Flask and its applications, structured similarly to "home.html." The key point here is that Flask allows developers to separate the structure and content of web pages, making it easier to manage and update the application's appearance. Additionally, Flask supports the creation of a parent template, as demonstrated by the "template.html" file. This template includes a navigation menu that can be shared across multiple pages, avoiding the need to duplicate code. Child templates, such as "home.html" and "about.html," extend the parent template, inheriting its structure and allowing for the insertion of content specific to each page. To enhance the styling of the web application, Cascading Style Sheets (CSS) are introduced. CSS files are stored in a "static" folder, separate from the "templates" folder. The "template.css" file is linked to the parent template, providing styles for the navigation menu, header, and other elements. This separation of concerns—HTML for structure, CSS for styling, and Flask for dynamic content—contributes to a cleaner and more maintainable codebase.

So, Flask serves as a powerful tool for web development in Python, enabling the creation of dynamic and visually appealing web applications. By handling HTTP requests and responses, defining routes, and seamlessly integrating with HTML templates, Flask simplifies the process of building and maintaining web applications. The use of templates and static files, along with the ability to create a parent template for shared elements, demonstrates Flask's flexibility and efficiency in developing web solutions.

3.3 Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put it in a formatted way. So, for this, we use data pre-processing tasks. Real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing requires tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

4. Results and discussion

Imports and Setup: The code imports necessary modules from the Flask framework. A Flask application is created and stored in the variable **app**.

Loading the Machine Learning Model: The script loads a machine learning model from a file (**DTC_Adaboost_model.pkl**) using the **pickle** module. This model is an Adaboosted Decision Tree Classifier (DTC).

Classification Function: There is a function called **classify_water_quality** that takes various water quality parameters (pH, hardness, solids, etc.) as input. This function utilizes the loaded machine learning model to predict the water quality based on the provided parameters.

Web Routes: Two routes are defined for the web application:

1. The default route **'/'** renders an HTML form (**water_quality_form.html**) where users can input water quality parameters.
2. The **'/classify'** route handles form submissions via POST request.

HTML Rendering:

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

- The **show_form** function renders the initial HTML form.
- The **classify** function processes the form data submitted by the user, converts the input to floats, and passes it to the **classify_water_quality** function.
- The classification result is then rendered in another HTML template (**result.html**).

Running the Application: The script checks if it's the main module (**__name__ == '__main__'**) and, if so, starts the Flask application in debug mode.

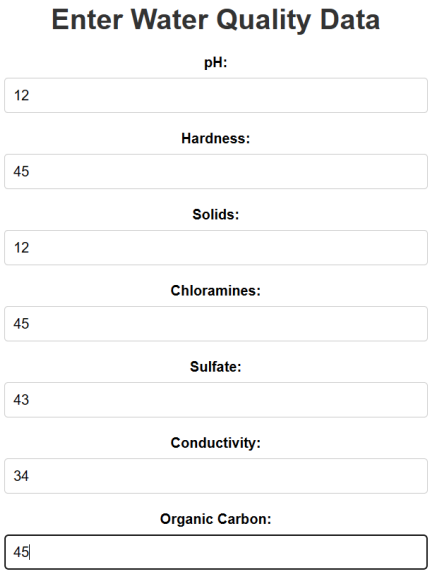

Figure 2. User interface of flask server.

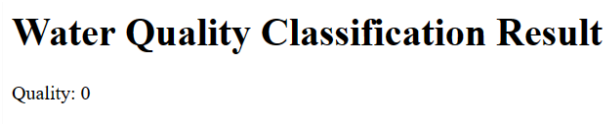

Figure 3. Prediction outcome.

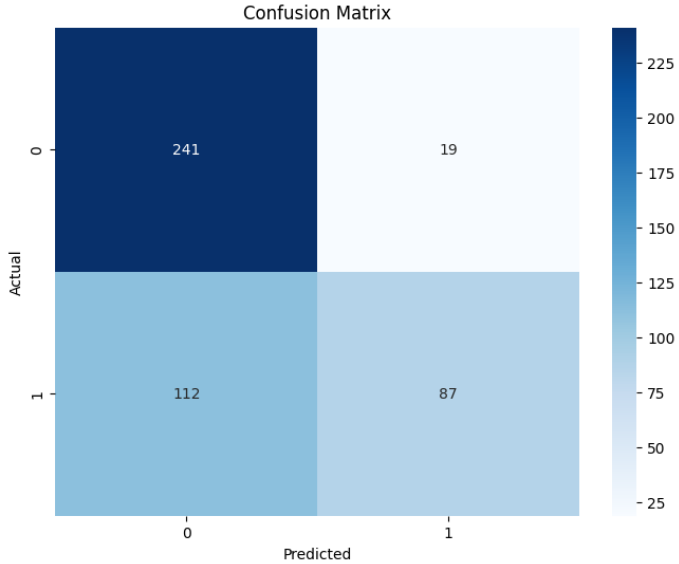

Figure 4. Existing KNN confusion matrix.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

	precision	recall	f1-score	support
0	0.98	0.99	0.98	260
1	0.99	0.97	0.98	199
accuracy			0.98	459
macro avg	0.98	0.98	0.98	459
weighted avg	0.98	0.98	0.98	459

Figure 5. Proposed DTC-Adaboost classification report.

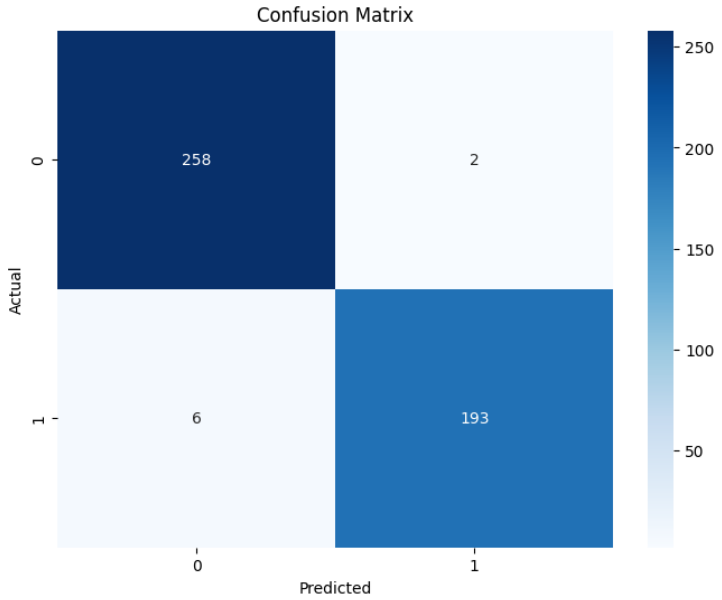

Figure 6. Proposed DTC-Adaboost confusion matrix.

5. Conclusion

In conclusion, the research journey undertaken for water potability classification, incorporating a Flask server, HTML-based UI, and binary classification models (KNN and DTC-Adaboost), represents a cohesive and innovative approach to addressing the critical need for efficient water quality assessment. The amalgamation of technological sophistication and user-friendly design underscores the project's commitment to not only technical accuracy but also practical applicability in diverse settings. The training of machine learning models on user-provided data enhances the system's adaptability and relevance in real-world scenarios, ensuring a dynamic and responsive classification mechanism. The performance estimation phase serves as a crucial benchmark, illuminating the effectiveness of the trained models in binary classification. The robust evaluation metrics employed provide a comprehensive understanding of the models' accuracy and reliability, guiding potential refinements for optimal performance. The subsequent data prediction from test data affirms the system's ability to generalize and make accurate predictions beyond the training dataset, substantiating its potential as a valuable tool in the realm of water quality assessment.

References

[1] World Health Organization. Guidelines for Drinking-Water Quality: First Addendum to the Fourth Edition; WHO: Geneva, Switzerland, 2017.

[2] Nouraki, A.; Alavi, M.; Golabi, M.; Albaji, M. Prediction of water quality parameters using machine learning models: A case study of the Karun River, Iran. Environ. Sci. Pollut. Res. 2021, 28, 57060–57072. [PubMed]
[3] UN Environment Programme. A Snapshot of the World's Water Quality: Towards a Global Assessment; United Nations Environment Programme: Nairobi, Kenya, 2016.
[4] Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. J. Environ. Chem. Eng. 2021, 9, 104599.
[5] Mishra, B.K.; Regmi, R.K.; Masago, Y.; Fukushi, K.; Kumar, P.; Saraswat, C. Assessment of Bagmati river pollution in Kathmandu Valley: Scenario-based modeling and analysis for sustainable urban development. Sustain. Water Qual. Ecol. 2017, 9, 67–77.
[6] Ewaid, S.H.; Abed, S.A. Water quality index for Al-Gharraf river, southern Iraq. Egypt. J. Aquat. Res. 2017, 43, 117–122.
[7] Ramakrishnaiah, C.; Sadashivaiah, C.; Ranganna, G. Assessment of water quality index for the groundwater in Tumkur Taluk, Karnataka State, India. E-J. Chem. 2009, 6, 523–530.
[8] Ewaid, S.H.; Abed, S.A. Water quality assessment of Al-Gharraf River, South of Iraq using multivariate statistical techniques. Al-Nahrain J. Sci. 2017, 20, 114–122.
[9] Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 2020, 585, 124670.
[10] Nearing, G.S.; Kratzert, F.; Sampson, A.K.; Pelissier, C.S.; Klotz, D.; Frame, J.M.; Prieto, C.; Gupta, H.V. What role does hydrological science play in the age of machine learning? Water Resour. Res. 2021, 57, e2020WR028091.
[11] Jafar, R. Assessment of surface water quality by using multivariate statistical techniques. Tishreen Univ. J. Eng. Sci. Ser. 2022, 44, 11–31.
[12] Ahmed, M.; Mumtaz, R.; Hassan Zaidi, S.M. Analysis of water quality indices and machine learning techniques for rating water pollution: A case study of Rawal Dam, Pakistan. Water Supply 2021, 21, 3225–3250.
[13] Bedi, S.; Samal, A.; Ray, C.; Snow, D. Comparative evaluation of machine learning models for groundwater quality assessment. Environ. Monit. Assess. 2020, 192, 776.
[14] Gupta, S.; Gupta, S.K. Evaluation of River Health Status Based on Water Quality Index and Multiple Linear Regression Analysis. In Sustainable Environmental Engineering and Sciences: Select Proceedings of SEES 2021; Springer: Berlin/Heidelberg, Germany, 2023; pp. 77–85.