IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 March 2024, accepted 31 March 2024, date of current version 10 April 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3384607*

RESEARCH ARTICLE

# Blockchain-Orchestrated Intelligent Water Treatment Plant Profiling Framework to Enhance Human Life Expectancy

**DHRUV SARJU THAKKAR[1], ANERI THAKKER[1], RAJESH GUPTA[1], (Member, IEEE), NILESH KUMAR JADAV[1], (Graduate Student Member, IEEE), SUDEEP TANWAR[1], (Senior Member, IEEE), GIOVANNI PAU[2], (Senior Member, IEEE), GULSHAN SHARMA[3], FAYEZ ALQAHTANI[4], AND AMR TOLBA[5], (Senior Member, IEEE)**

[1]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat 382481, India
[2]Faculty of Engineering and Architecture, Kore University of Enna, 94100 Enna, Italy
[3]Department of Electrical Engineering Technology, University of Johannesburg, Johannesburg 2006, South Africa
[4]Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia
[5]Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia

Corresponding authors: Giovanni Pau (giovanni.pau@unikore.it), Sudeep Tanwar (sudeep.tanwar@nirmauni.ac.in), and Rajesh Gupta (rajesh.gupta@nirmauni.ac.in)

This work was supported in part by King Saud University, Riyadh, Saudi Arabia, through the Researchers Supporting Project under Grant RSPD2024R681; and in part by the Kore University of Enna, Enna, Italy.

**ABSTRACT** Water quality degradation has turned out to be of crucial importance due to various factors over the past decade. Pollution, climate change, and population growth are the factors that affect water quality. Contaminations such as microorganisms, heavy metals, and excessive nitrogen and phosphorous disrupt water pH levels, posing significant health risks. Despite the innovation in the Internet of Things(IoT), allowing balancing the pH by adding chlorine and fluoride after the disinfection step, several security issues(e.g., distributed denial of service, data manipulation, and session hijacking) manoeuvre the operational performance of the water treatment plants. This causes people to consume polluted water, which has many adverse effects on human health and reduces life expectancy. To address this critical concern, we propose a novel approach integrating artificial intelligence(AI) and blockchain technology into water treatment plant management. Our methodology utilizes a standard water quality dataset, which has features such as pH and total hardness, which is used for binary classification, indicating water as potable or not potable. We employ various AI classifiers such as stochastic gradient descent classifier (SGDC), decision tree (DT), Naive Bayes (NB), K nearest neighbours (KNN), and logistic regression (LR). Furthermore, an InterPlanetary File System(IPFS)-based public blockchain is integrated to resist the data manipulation attack, where the potable water sample is securely stored in the blockchain's immutable ledger. The proposed model is evaluated using various performance metrics such as confusion matrix analysis, learning curve assessment, training accuracy, and blockchain scalability. Notably, the DT model emerges as the best-performing classifier with an accuracy of 99.41% and scalability of 35 with 120 data transactions.

**INDEX TERMS** Artificial intelligence, water treatment plants, water profiling, blockchain, Internet of Things (IoT), IPFS.

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

## I. INTRODUCTION

Water is an essential need for all living beings. Apart from drinking, household, cooking, and industrial water usage,

drinking is the most essential. It is crucial for the survival of human beings. Water not only quenches your thirst but is also an important functional unit of life. About 60-70% of the human body is made up of water, which regulates body temperature and moistens tissues of body organs. It helps in blood circulation as it can carry nutrients and oxygen. It can lessen the burden on kidneys, livers, and other parts of the human body and also removes toxins from our bodies. These benefits exist only if the quality of water we drink is good; otherwise, it causes diseases, which signifies the crucial role of water potability in improving life expectancy. The Composite Water Management Index Report 2018 of the National Institution for Transforming India Aayog stated that about two lakh people die every year due to inadequate access to safe water, which shows the adverse effects of impure water on human life expectancy. NITI Aayog (India) estimated that by 2030, around 600+ million people will face water stress, which is almost 40% of India's population [1].

The paper [2] explored the challenges of understanding and managing the potential impacts of climate change on India's water resources. Human activities, such as those in the energy, land, water, and climate nexus, complicate these assessments, as they can either amplify or mitigate climate change effects. To develop effective policies, a holistic approach is crucial, recognizing these interconnections and differentiating between local human actions and global climate impacts.

The [3] highlights the importance of access to clean water, sanitation, and hygiene (WASH) for human health and well-being. It notes that chemical pollution poses a significant threat and that contaminated water can lead to various health problems, particularly diarrheal diseases. Over a million deaths from diarrhoea occur annually due to poor WASH conditions, and these conditions also hinder the prevention and treatment of malnutrition and neglected tropical diseases. However, studies demonstrate that improving WASH services can drastically reduce diarrheal deaths, emphasizing the urgent need for interventions to address these challenges. A study [4] found serious problems with the quality of drinking water in Thulamela, South Africa. While water from the source (dam) was safe, water from taps and storage containers in homes had bacteria. This shows that contamination is happening while water is being moved and stored. The study [4] found that many households don't treat their water before drinking it, which increases the risk of getting sick from waterborne diseases. Water shortages also increase the risk of contamination because the water can become stagnant and less protected. The research also suggested that lead in the water may be linked to cancer, which highlights the importance of tighter regulations and better monitoring of trace metals in our drinking water. Hence it is crucial to tackle water quality issues because they pose a serious risk to health and lifespan. Contaminated water can spread diseases and cause multiple health problems, endangering the well-being of individuals.

As per the review [5], non-portable water is a major cause of mortality globally. A WHO report revealed that poor drinking water quality contributes to over 50% of diseases and alarmingly accounts for 80% of global illnesses. Its impact is particularly devastating for children, as it is responsible for half of childhood deaths worldwide. Research [6] has also connected swimming in contaminated water to a higher risk of illness, with children under the age of ten showing the highest rates of infection. Furthermore, the use of drinking water contaminated with arsenic has been connected to bladder, kidney, and skin cancers as per the article [7]. The author of [8] deduced that the chance of developing various types of cancer is increased by carcinogens added during water treatment procedures, such as chlorine treatment.

Water quality is affected by many factors, including pH, minerals (sodium, magnesium, calcium, chloride, potassium, carbonate, sulfate), total dissolved solids (TDS), electrical conductivity (EC), total hardness (TH), and the Water Quality Index (WQI). These factors influence whether water is safe to drink and affect human health. Unbalanced levels of these factors can cause digestive issues, kidney problems, heart disease, and brain disorders. Improving water quality is crucial for public health and life expectancy by providing access to clean, safe drinking water. Altering water quality data can dramatically harm human health. Inaccuracies in reporting can cause consumption of contaminated water, leading to stomach problems, lack of hydration, and serious long-term health issues. In areas with limited water supply, these problems become even more concerning, especially in cities with growing populations and pollution issues. To address this, it's crucial to accurately monitor water quality and implement sustainable practices that guarantee access to clean drinking water for everyone.

Water pollutants pose serious threats to human well-being, the natural environment, and the variety of life on Earth. The authors of [9] drew attention to the study that used polyethene particles as a typical type of microplastics and analyzed their behaviour in comparison to dissolved substances, using energy loss, residence time distribution, and mixing models, which emphasize the need for understanding how microplastics move through city water systems.ies like carbon nanotubes and graphite oxides. Emphasis is placed on in-depth studies that consider not only the original pollutants but also their byproducts. These efforts encompass identifying, measuring, and evaluating the risks associated with these contaminants.

The researchers of [10] examined pollution in drinking water caused by nitrate contamination and its possible health risks. The presence of nitrate in water poses a significant obstacle to ensuring clean and safe drinking water, with the World Health Organization (WHO) [11] setting a safety limit of 50 mg/L. Water samples were collected from different areas of Tehran during wet and dry seasons. Nitrate levels were higher in the dry season because of groundwater with

high nitrogen levels. The study in the paper [10] showed that nitrate contamination in Tehran's drinking water poses a high cancer risk, especially for children. Over half the samples collected during the dry season showed a risk of cancer for different groups of people, even though only a few samples had nitrate levels above the safety standards for drinking water.

The study in the paper [12] examined the movement of microplastics in sewer systems, specifically in manholes, which are crucial junctions in these networks. The study used polyethene particles as a typical type of microplastics and analyzed their behaviour in comparison to dissolved substances, using energy loss, residence time distribution, and mixing models, which emphasize the need for understanding how microplastics move through city water systems. This knowledge is vital for reducing the risks of contamination and improving the design and management of wastewater infrastructure. The results of ref. [12] showed that most of the plastic particles move through the system like dissolved substances, but a portion gets trapped in manholes with high water levels. The longer these particles stay trapped, the greater the risk of contamination.

The quality of water also depends on the sources from where the water comes. Usually, the water sources include ground and surface water. The water from these resources must be treated first before being used for drinking. Preserving water quality is vital, especially as climate change intensifies water scarcity. Ensuring the quality of accessible water is crucial for human health, environmental stability, and economic sustainability, especially as water resources become scarcer. Climate change can affect water quality by altering precipitation patterns, disrupting hydrological cycles, and increasing extreme weather events. The study in the paper [13] showed that unsustainable practices and climate change are the main causes of the notable statewide drop in groundwater recharge. According to the authors of [13], more than 80 million people might not have access to water as a result of this reduction, highlighting the urgent need for innovative management strategies and policies to lessen the effects of diminishing water supplies. The results highlight the need for precise monitoring and long-term management plans to maintain water quality in the face of rising hydroclimatic swings and intensifying water shortage issues. The Water Treatment Plant (WTP) collects water from different resources and uses different methods such as coagulation/flocculation, sedimentation, filtration, disinfection, sludge drying, fluoridation, and pH correction. To perform all these methods, the WTP is classified into three main categories: a pretreatment plant, a post-treatment or demineralized water plant, and a waste treatment or effluent treatment plant.

The authors of [14] proposed a process for getting clean water using coagulant from locally abundant kaolin clays. Alum is a coagulant that is used for raw water treatment plants. They identified that a dosage of 40mg/L of the extracted alum showed effective coagulant properties with great potential to treat raw water. Ref. [15] proposed a lab-scale experiment with a coconut-shell-based granular activated carbon column that showed the possibly different mechanisms of removal between perfluorohexyl sulfonate and perfluorooctanoic acid, indicating that the sulfonate-based perfluorinated compounds may be a limiting factor in the granular activated carbon replacement cycle for perfluorinated compounds removal.

Numerical modelling plays a vital role in optimizing water treatment plants. These models simulate various scenarios and conditions, enabling engineers to predict potential issues and refine the treatment processes. By accurately estimating water flow, energy consumption, and pollutant removal efficiency, models guide resource allocation and operational decisions. Furthermore, by forecasting the impact of operational changes on water quality and offering insights into the facility's performance, numerical modelling helps ensure compliance with regulatory standards.

Researchers [16] developed an improved turbine design for Micro Hydroelectric Plants. The design focused on extracting energy from Wastewater Treatment Plants, which have a lot of unused low-energy water flow. The design used numerical modelling and a special algorithm to refine the shape of the turbine blades and other parts. The result is a propeller turbine that can extract up to 76% of the available energy, making it a promising option for generating renewable energy from WWTPs.

The study [17] used numerical modelling to investigate how water temperature impacts disinfection in tanks that use chlorine. The simulation considers the movement of substances, chlorine breakdown, and how temperature affects water's density and thickness. The results of the study [17] showed that even small temperature changes can significantly alter the amount of chlorine in the water and the effectiveness of the disinfection process, potentially affecting water treatment operations. The model's accuracy had been rigorously tested against real-world data, demonstrating its reliability in optimizing temperature settings for efficient water treatment and reducing the creation of disinfection byproducts.

By examining the physical-chemical hydrodynamics of activated sludge reactors and their effect on N2O emissions, this research [18] seeks to bridge knowledge gaps in the design of these reactors. The paper examined variables influencing the accuracy of aeration and N2O emission predictions in surface-aerated oxidation ditch-type reactors using computational fluid dynamic simulations. The study [18] emphasized the importance of design, operational circumstances, and biokinetic factors in forecasting N2O emission by comparing laboratory results with actual observations in a water treatment facility.

Numerical modelling in water treatment lacks automation, leading to a labour-intensive and time-consuming process. It involves manual steps at every stage, such as setting up the model, entering data, and interpreting findings. The smart

sensors employed in the WTP ecosystem apply numerical approaches that will significantly raise the computations, and eventually, the solutions become computationally expensive. Since it also relies on human interaction, mistakes and biases can creep in, compromising the precision and dependability of the results. WTP could, therefore, find it difficult to maximize their performance and effectively adjust to changing circumstances.

Traditional water treatment facilities struggle with lengthy processes, expensive operations, poor monitoring systems, obsolete equipment, and unreliable record-keeping. These problems hinder efficiency, compromise water quality, and pose potential health hazards. However, Artificial Intelligence(AI) offers solutions that automate tasks, enabling a more efficient and adaptable approach to modeling and optimizing processes. AI can improve water treatment plants by making them more efficient. AI systems can automate monitoring, maintaining equipment, and keeping records. This frees up human workers and improves plant performance. AI algorithms can also analyze large amounts of data to optimize water purification processes, making them more effective, less expensive, and faster. By harnessing AI, water treatment plants can deliver clean, safe water to communities more efficiently.

The authors of [19] proposed an IoT and machine learning-based solution to have an automatic, cost-effective WTP. The proposed system used Thing Speak to monitor all the data using the ESP8266 Wi-fi module. Biochemical oxygen and chemical oxygen demand sensors have been integrated to measure the quality of water accurately. Then, the ML algorithm is applied to the sensor information to extend the exactness of the sensor information. The study [20] used machine learning techniques like Support Vector Regression and Regression Trees to estimate wastewater quality indicators from urban catchment data, which can be helpful for wastewater management. Support Vector Regression performed better than Regression Trees in predicting total suspended solids, total dissolved solids, and chemical oxygen demand, while both models were similar in predicting biochemical oxygen demand. These results showed that machine learning could be useful for planning wastewater treatment plants, managing wastewater in real-time, and addressing environmental challenges related to wastewater.

The articles [19] and [20] explored the application of AI and machine learning algorithms in WTPs to monitor and assess water quality. However, there are higher chances of security attacks in the approaches mentioned in the papers [20] and [19]. So in this scenario, security is the biggest concern. To overcome the problem of data security, we proposed an AI and blockchain-based approach. In this paper, we have used various ML algorithms such as the Stochastic Gradient Descent classifier (SGDC), Decision Tree (DT), Naive Bayes (NB), K Nearest Neighbours (KNN), and Logistic Regression (LR) to classify the water sample into drinkable and non-drinkable water. Also, to confront the data manipulation attack and data storage cost, we integrated the InterPlanetary File System (IPFS) protocol to store the water-related data securely.

### A. MOTIVATION

The motivation for the proposed approach is as follows.

- The profiling of water is a necessary aspect of streamlining operations and providing clean water. To make this process simpler, more accurate, and safer, multiple approaches have been researched and deployed, such as numerical modelling, IoT-based solutions, and automated solutions using AI.
- In the existing state-of-art works, References [16], [17], and [18] have explored the numerical modelling approach for the water treatment applications which lacks automation in the applications that may lead to human errors and biases. Moreover, smart sensors are resource-constrained devices; applying numerical optimization on such resource-constrained devices can degrade the operational performance of the smart grid environment. This motivated us to search for methodologies that are more accurate and automated, like AI.
- The approach used by [19] and [20] discovered the application of machine learning techniques for forecasting water quality. Their approach helps with accurate results and with minimal human biases and errors. Still, their approach fails to secure the predicted data from data tampering attacks. This led us to leverage machine learning techniques blended with a blockchain-based approach to make accurate results, keeping data secure and immutable.
- Furthermore, the aforementioned approaches rely on a single technology, such as either AI or blockchain, to offer security enhancement in WTP. We want to state that amalgamating two technologies, i.e., AI + blockchain, can substantially strengthen the security of WTP. In addition, the existing works related to blockchain have not adopted the essential benefits of IPFS in their proposed solutions, thereby increasing the latency and reducing the scalability of their solutions.

*Novelty:* Limited literature exists for securing WTP from adversaries aiming to deteriorate the operational performance of smart sensors and contaminate the water supply with hazardous chemicals or pollutants, thereby endangering the health of the nation's citizens. Already available literature relied on simple AI modules or blockchain technology to offer security in WTP. Nevertheless, they have not blended two technologies, such as AI and blockchain, to provide more robust and resilient security features in the WTP profiling systems. In the proposed work, we used AI models to solve a binary classification problem wherein the class labels are potable and not potable water data. In the non-potable water data, the attackers have manipulated the smart sensors so that they can release high amounts of chemical compounds in the

treatment plants. Consuming such high chemical water can reduce human life expectancy.

Hence, the utilized AI model in the proposed work efficiently classifies the water potability data based on each feature's threshold. Further, there is a severe possibility that the attackers can perform data manipulation attacks on the predicted data (from AI models), thereby raising data integrity concerns. To respond to this challenge, we adopted an IPFS-based blockchain technology that provides secure data storage for the predicted data in the blockchain's immutable ledger. Integrating IPFS in the blockchain improves the response time and thus increases the scalability of the proposed solution. DT model emerges as the best-performing classifier, which classifies the portable and non-portable water with an accuracy of 99.41%, and blockchain offers scalability of 35 with 120 data transactions. The existing works have not considered the aforementioned approaches and staggering benefits of AI and blockchain amalgamation; thus, their solutions are not end-to-end secure and fail to demonstrate the full potential of safeguarding WTPs from adversaries aiming to compromise operational efficiency and endanger public health with contaminated water.

### B. RESEARCH CONTRIBUTIONS

The following are the research contributions of this paper:

- We proposed a blockchain and AI-based water profiling framework that classifies water into potable and non-potable water to increase human life expectancy.
- Further, to secure the predicted data from data tampering attacks, we adopted IPFS-based blockchain technology that offers secure data storage to the predicted potable data.
- The proposed framework is assessed by considering different evaluation metrics, such as training time, log-loss score, training accuracy, blockchain response time, and scalability.

We have organized this paper as follows. Section II gives an overview of the existing schemes for profiling the water from different water treatment plants. Section III describes a module for predicting the potability of water for enhancing human life expectancy. Section IV illustrates the proposed framework using an AI layer and a blockchain layer. Section V discusses about the results we obtained. Section VI concludes the paper.

## II. RELATED WORKS

A lot of research has been carried out on water treatment by researchers to help determine whether drinking water is safe. Across the globe, researchers used different techniques and algorithms to classify drinking water. The authors of [29] proposed the hydroclimatic modelling approach for recognizing and reducing how climate change affects water supplies and related processes. Hydroclimatic modelling investigates how climate variations affect water resources and events like floods. To handle these issues, the research introduces a new method that combines surrogate modelling and physics-based machine learning. Surrogate modelling uses simplified yet effective models to mimic complex systems, improving computational efficiency. The authors proposed a Physics-informed neural network-based surrogate model that incorporates physical principles into the neural network structure. This integration enhances the model's ability to capture crucial physical behaviours, ensuring greater accuracy and interpretability in hydroclimatic simulations. Using this approach in flood simulators based on Shallow Water Equations, the study shows substantial enhancements in predicting floods better than existing data-driven methods. This progress is critical in hydroclimatic modelling, where accurate flood prediction is vital for comprehending and minimizing the effects of climate change on water sources and connected systems.

Researchers [30] developed a numerical model to study a dam-controlled river system, balancing the need for water for human use and the health of the river ecosystem. They used the model in a section of China's Jinjiang watershed. The performance of the model was tested by comparing predicted water levels and water quality measurements (Chemical Oxygen Demand and $NH_3$ -N) with real-life data. The model's predictions were generally close to the actual measurements, with relative RMSE ranging from 5.5% to 28.4%. The model was then used to develop plans for managing dam operations during drought conditions and water contamination events. In the dry scenario, the dam release plan was successful in meeting minimum water flow requirements for the environment. However, in the contamination scenario, the plan was only able to limit the spread of pollution, but it did not fully meet water quality standards.

The authors of [31] proposed mathematical frameworks to optimize the design of water treatment systems. The need for water is growing due to population growth and climate change. The first model, Mixed Integer Nonlinear Programming, aimed to minimize the water system's cost. However, it can be unstable due to its complex nature. To address this, two enhanced models were introduced: a partially linearized Mixed Integer Nonlinear Programming and a Mixed Integer Linear Fractional Programming model. These models were more stable and could handle the complexities of water treatment processes. The models were tested on case studies involving seawater desalination and surface water treatment for drinking water production. The mixed integer linear fractional programming model outperforms other methods of designing water treatment systems. It provides better solutions in less time. This approach aligns with industry standards. The study shows that using optimization techniques improves the efficiency of designing water treatment processes.

The research [32] presented a control system for wastewater plants using an optimization algorithm called a Non-dominated sorting-based multi-objective cuckoo search optimization algorithm. This algorithm optimized the parameters

**TABLE 1. Comparative analysis of the proposed work with existing works.**

| Author | Year | Objective | Methodology | Pros | Cons | Performance Analysis |
|---|---|---|---|---|---|---|
| Safder *et al.* [21] | 2022 | effluent concentration of total nitrogen prediction | multi head-attention-based gated recurrent unit, partial least squares, multiple linear regression, multilayer perceptron, LSTM, gated recurrent unit | presented a prediction model used to know the concentration of total nitrogen a few hours ahead | Didn't discuss the effect of the concentration of total nitrogen present in water on human health and human life expectancy | Multihead-attention-based gated recurrent unit: 98% accuracy rate |
| Patel *et al.* [22] | 2022 | potability prediction | RF classifier, XGBoost, Decision Tree, AdaBoost, SVC | proposed a prediction model to predict the water potability | Didn't discuss the effect of water potability on human life expectancy and security aspects | RF classifier results: 81% accuracy. |
| Alqahtani *et al.* [23] | 2022 | river water salinity prediction in terms of electrical conductivity and dissolved solids | RF, gene expression programming, ANN | proposed RF model approach gives better results to predict river water salinity in terms of electrical conductivity and dissolved solids | Didn't consider the effect of electrical conductivity and dissolved solids on human health | RF model results: For total dissolved solids, while testing: NSE=0.91, R2 = 0.93, RMSE = 3.1, MAE = 5.10. For electrical conductivity, while testing: NSE = 0.930.91, R2 = 0.93, RMSE = 3.52, MAE = 2.5. |
| Jalal *et al.* [24] | 2019 | water quality monitoring | DT, SVM | presented the performance analysis of the famous classification algorithms on the data collected from a Tunisian water treatment station | Didn't discuss the security aspects related to data | SVM: 98% accuracy with 1/4 samples. |
| Wongburi *et al.* [25] | 2022 | sludge volume index prediction | recurrent neural networks, explainable AI | prediction of sludge volume index help to establish corrective measures for maintaining the stable value of and hence improve operational performance | Didn't consider data security and human health expectancy based on the predicted sludge volume index | For the first dataset(data from 1996-2000), RMSE is 4.161 and MAE is 3.284, while for the second(data from 2001-2020) and third(data from 2010-2020) datasets, the values are lower 3.360 and 2.156. |
| Manimekalai *et al.* [26] | 2022 | increase the degree of accuracy that can be achieved by wastewater treatment models | ANN-ELM, chemical oxygen demand model, biological oxygen demand model, deep convolutional neural network- sine cosine algorithm model | presented approach of estimation of the concentration of chemical oxygen demand in the effluent and the input of wastewater treatment systems | Haven't taken the effect of chemical oxygen demand and other inputs on human life expectancy | ANN-ELM: 94.17% accuracy and 30.66% RMSE. |
| Banerjee *et al.* [27] | 2022 | contamination of water prediction | multivariable linear regression, support vector regression, decision tree regression, lasso regression | proposed an approach for the prediction of contamination of water considering the factors such as water body location, latitude, longitude, and elevation | Didn't discuss the effect of water-related parameters on human life expectancy and security aspects | Multivariable linear regression excelled for temperature, biological oxygen demand, and dissolved oxygen. Support vector regression outperformed for pH and chlorides parameters. |
| R Kshirsagar *et al.* [28] | 2022 | effluent chemical oxygen demand prediction | ANN | presented a prediction model to predict the input and effluent chemical oxygen demand for effluent treatment procedures which enhances the performance of wastewater treatment plant | Didn't consider the effect of chemical oxygen demand on water potability | Two ANN-based techniques: accuracies of 92.67% and 89.23% in predicting influential and effluent chemical oxygen demand. |
| Donnelly *et al.* [29] | 2024 | flood prediction and hydroclimatic modeling for minimizing the impact of climate change on water sources and related systems | Physics-informed neural network-based surrogate model | proposed solution is highly efficient, accurate, versatile and robust. | Didn't discuss the water quality parameters. The proposed solution lacks scalability. | reduces the error rate by around 11%. |
| Gao *et al.* [30] | 2023 | to study a dam-controlled river system and balance the need for water to plan for drought scenarios | hydrodynamic model and water quality models | proposed model uses to develop plans for managing dam operations during drought conditions and water contamination events | The proposed solution has less reliability, poor extensibility, and which could lead to ecological problems. | The statistical model has RRMSE values of 23.6% for COD and 28.4% for NH3-N. |
| Koleva *et al.* [31] | 2017 | to optimize the design of water treatment system | a partially linearized Mixed Integer Nonlinear Programming and a Mixed Integer Linear Fractional Programming model | The proposed approach aligns with industry standards and optimization techniques to improve the efficiency of designing water treatment processes | Didn't discuss the effect of water portability on human life expectancy, security, and scalability aspects | The Mixed Integer Linear Fractional Programming model achieves a production rate of approximately 600,000 m3/d, while exhibiting a cost reduction of 60% with shorter computational times. |
| Aparna *et al.* [32] | 2023 | to present a control system that improves effluent quality, reduces operating costs of treatment plants | Non-dominated sorting-based multi-objective cuckoo search optimization algorithm. | the study helps optimize wastewater treatment and reduces pollution levels and flow instability in all weather conditions | Didn't discuss other parameters regarding water treatment other than dissolved oxygen. | The proposed approach reduces pollution levels ranging from 0.63% to 1.3%. |
| The proposed framework | 2023 | water profiling to enhance human life expectancy | LR, NB, KNN, SGDC, DT | Accurate, efficient and secure scheme for enhancing human life expectancy based on water potability | - | Decision tree model: accuracy of 99.41% and scalability of 35 with 120 data transactions. |

of a controller used to regulate the amount of oxygen in the water, which is crucial for efficient wastewater treatment. The system considered two goals: one focusing on improving effluent quality and reducing operating costs, and the other on reducing nitrogen and ammonia levels. By using the model given by [32] to find the best parameters for the controller,

the study helps optimize wastewater treatment. Simulations demonstrate that their approach reduces pollution levels and flow instability in all weather conditions, with reductions ranging from 0.63% to 1.3%.

The authors in [21] proposed an intelligent scheme for wastewater treatment plants using AI and deep learning models for predicting the effluent concentration of total nitrogen a few hours ahead. They proposed a multistep-ahead effluent total nitrogen prediction framework at wastewater treatment plants under dynamic variational data using several ML and recurrent neural network models, including partial least squares regression, multiple linear regression, multilayer perceptron, long-short term memory, gated recurrent unit, and multihead-attention-based gated recurrent unit. The multihead-attention-based gated recurrent unit method reported the most accurate and selective predictive performance among all introduced models in predicting the total nitrogen present in the water.

The authors of [22] proposed an ML-based water potability prediction model by using the synthetic minority oversampling technique and explainable AI, which showed the comparative analysis of different machine learning approaches like support vector machine (SVM), DT, random forest (RF), gradient boost, and ada boost, used for the water quality classification. In this analysis, they got the highest accuracy of 0.80 using the RF classifier. Later, [23] presented the comparison of individual supervised ML models, such as gene expression programming and artificial neural network, with that of an ensemble learning model, i.e., RF, for predicting river water salinity in terms of electrical conductivity and dissolved solids. In terms of accuracy, the RF model outperforms other models on the training and testing datasets, followed by gene expression programming and artificial neural network models. The highest R2 value and lowest root mean square error value were both attained by the RF model.

The authors of [24] presented a water quality monitoring system using machine learning techniques. The algorithm used was DT and SVM. The linear SVM, Fine tree, and Medium tree for the full samples give similar precision. On the other hand, for $\frac{1}{4}$ of the samples, the linear SVM gives better precision. Then, Wongburi and Park [25] proposed an intelligent scheme for wastewater treatment plants using recurrent neural networks and explainable AI to predict sludge volume index over the time series data. They collected data from 1996 to 2020. After creating the appropriate datasets and training the datasets, the prediction model had a root mean square error of 4.161 and a mean absolute error of 3.284 for the first dataset (data from 1996 to 2000). The second dataset(data from 2001 to 2020) has a root mean square error of 3.360 and a mean absolute error of 2.156, which is analogous to the third dataset (data from 2010 to 2020).

The authors of [26] proposed an overview of artificial neural network models that have been developed over the past two and a half decades for membrane procedures that are used in the management of wastewater and the purification of water. The artificial neural network-extreme learning machine (ANN-ELM) model, chemical oxygen demand, and biological oxygen demand models each have an accuracy value of 82.17%, 85.68%, and 88.31%, respectively, while the deep convolutional neural network- sine cosine algorithm has an accuracy value of 90.11% for 100 data points from the dataset. In comparison, the suggested ANN-ELM approach has an accuracy value of 93.14%. Although the ANN-ELM model did the best, with an accuracy of 94.17%, the accuracy for the ANN-ELM, chemical oxygen demand, and biological oxygen demand models, as well as the accuracy for the deep convolutional neural network- sine cosine algorithm model, is, respectively, 83.77%, 86.82%, 89.24%, and 91.77% for 300 data points from the dataset. The ANN-ELM model has shown the maximum performance with less root mean square error for 600 data points at 30.66%, while the ANN-ELM, chemical oxygen demand, biological oxygen demand, and deep convolutional neural network- sine cosine algorithm models have root mean square error of 44.82%, 40.74%, 37.18%, and 32.85%, respectively.

The authors of [27] proposed a machine learning approach for the prediction of contamination of water considering factors such as water body location, latitude, longitude, and elevation. They used multivariable linear regression, support vector regression, decision tree regression, and lasso regression. For temperature, multi-variable linear regression showed the best result with an $R^2$ score of 40.94 in training and 52.51 in testing. For pH, support vector regression outperformed with an $R^2$ score of 92.09 in training and 46.97 in testing. For biological oxygen demand, multivariable linear regression demonstrated the best results with $R^2$ score of 99.99 in training and 99.99 in testing. For dissolved oxygen, multivariable linear regression reported the best results with an $R^2$ score of 69.58 in training and 54.83 in testing. For turbidity, hardness, and alkalinity, the authors were unable to conclude which model is better because of the low correlation among the data. For chlorides, support vector regression showed the best results with an $R^2$ score of 33.53 in training and 71.188 in testing. For chemical oxygen demand, multivariable linear regression outperformed with an $R^2$ score of 12.01 in training and 9.02 in testing. Ref. [28] presented the use of an artificial neural network (ANN) algorithm to enhance the performance of wastewater treatment plants. Their proposed system predicted the influential and effluent chemical oxygen demand for effluent treatment procedures with an accuracy of 92.67% with one model and 89.23% with another model.,

The researchers [29], [30], [31], and [32] explored the mathematical approach for maintaining water treatment processes. The smart sensors are resource-constrained devices in the WTP ecosystem; applying numerical-based approaches significantly raised the computations, and eventually, the solution became computationally expensive. With that deliberation, we focused on ML-based approaches in WTP. The authors [21], [23], [25], [28] incorporated ML into their

solution. In some of their research work, the authors have not focused on the water potability prediction and their effects on human life; rather, they have discussed how different components of water affect the quality of water, such as authors of [21] predicted the effluent concentration of total nitrogen a few hours ahead, authors of [28] predicted the input and effluent chemical oxygen demand, authors of [25] predicted sludge volume index and authors of [23] predicted water salinity. However, some of the researchers have not considered the effects of poor-quality water on human life expectancy and security aspect. Table 1 shows the comparative analysis of the existing work with the proposed framework.

## III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we describe a system model that predicts the potability of water from the WTP for enhancing human life expectancy. The system comprises different water treatment plants $\{w_1, w_2, w_3, \ldots, w_m\} \in W$. Here, each $w_i$ is equipped with sensors, such as the potential of hydrogen (pH) analyzer, conductivity analyzer, oxidation-reduction potential analyzer, and dissolved oxygen analyzer, which are represented as $\{s_1, s_2, s_3, \ldots, s_n\} \in S$. Each $s_i$ collects water-related data $\{\delta_1, \delta_2, \delta_3, \ldots, \delta_n\} \in \Delta$ from $w_i$ and sends to a real-time water profiling application $F$.

$$\begin{aligned} & w_i(\, s_j) \xrightarrow[\delta_j]{\text{sends}} F \\ & \forall i \; i \in \{1, 2, \ldots, m\} \\ & \forall j \; j \in \{1, 2, \ldots, n\} \end{aligned} \tag{1}$$

where $i$ denotes Water profiling applications use different methods to check the potability of water, whether the water is drinkable or not.

$$F = \begin{cases} \delta_i, & \text{drinkable} \\ \delta_i', & \text{not drinkable} \end{cases} \tag{2}$$

AI models are used to predict the potability of water from WTP to maximize human life expectancy. However, some attackers $\mathcal{K} \in \{k_1, k_2, \ldots, k_k\}$ can manipulate $\delta_i$ collected from $s_i$, which will lead to change in predicted results.

$$K\ (\delta_i) \xrightarrow[\text{sends}]{s_i} \text{Performance} \downarrow \delta_i' \notin \Delta \tag{3}$$

Therefore, there is a need for blockchain networks to restrict such data manipulation attacks. In the blockchain network, the data is stored in blocks $\{b_1, b_2, b_3, \ldots, b_m\} \in Blockchain_{\Delta}$. We considered $A_i$ as the max accuracy of the $w_i$ plant. X denotes the summation of $A_i$ obtained from each $w_i$. We inferred $\omega$ as the human life expectancy parameter, which is based on water potability, and $X_{thr}$ is the threshold value of the accuracy of WTP.

$$X \uparrow \implies \omega = \text{high} \tag{4}$$

$$X \downarrow \implies \omega = \text{low} \tag{5}$$

$$A_i = max_{\text{accuracy}} \left( \sum_{j=1}^{n} (\delta_j) \right) \tag{6}$$

## IV. PROPOSED FRAMEWORK

FIGURE 1 shows the proposed blockchain and AI-based water profiling framework for enhancing human life expectancy. The proposed framework is divided into two layers WTP and AI layers. The description of each layer is given as follows.

### A. WATER TREATMENT PLANT LAYER

In this layer, we integrate an AI-based classification algorithm to maximize human life expectancy by predicting the potability of water in the WTP. The data is collected using different sensors designed to find the nutritional content in the water. In this layer, we have different WTPs, each having a set of sensors installed that collect data on the nutritional content of the water. Therefore in each plant, we have a set of sensors $s_i$, where $\{s_1, s_2, s_3, \ldots, s_n\} \in S$. Each $s_i$ collects water-related data $\{\delta_1, \delta_2, \delta_3, \ldots, \delta_n\} \in \Delta$. This data represents various parameters and characteristics of water, such as pH level, conductivity analyzer, oxidation-reduction potential analyzer, dissolved oxygen analyzer and many more. We have also incorporated blockchain technology, which ensures the security and integrity of the data and that the data remains unaltered throughout its lifetime.

### B. AI LAYER

#### 1) DATASET DESCRIPTION

The data has been taken from IEEE data port [33]. Once the data is collected from the sensors in the water treatment plants, it is integrated and stored in comma-separated-values (CSV) file format. This process involves merging data from different sensors and making a unified dataset, which contains comprehensive information about the water in each water treatment plant. This integrated data is then organized into a tabular format, where each row corresponds to a specific data entry, and each column represents different characteristics of water. Finally, the data is stored in CSV file format, which provides a structured and easily accessible format for further analysis.These features serve as the basis for predicting the potability of water, with various ML techniques utilized to analyze and classify the data. The potability of water is predicted based on these features, as listed in Table 2. For the analysis of the dataset, we used a box plot to visually inspect the distribution of values for each of the 11 numeric features. FIGURE 2 shows the box plots of all the features. It makes it easier to interpret the data and detect outliers. The box plot is plotted on scaled data because the range of all the data varies in terms of magnitude. This helps compare the number of outliers and measures such as the median and quartiles. The use of box plots proved to be a valuable exploratory tool in our data analysis.

#### 2) DATA PREPROCESSING

To prepare the collected data and gain insights into its characteristics Exploratory Data Analysis (EDA) techniques are employed. EDA involves various methods to understand the dataset's structure, patterns and relationships. EDA helps us visualize the data in the form we want.

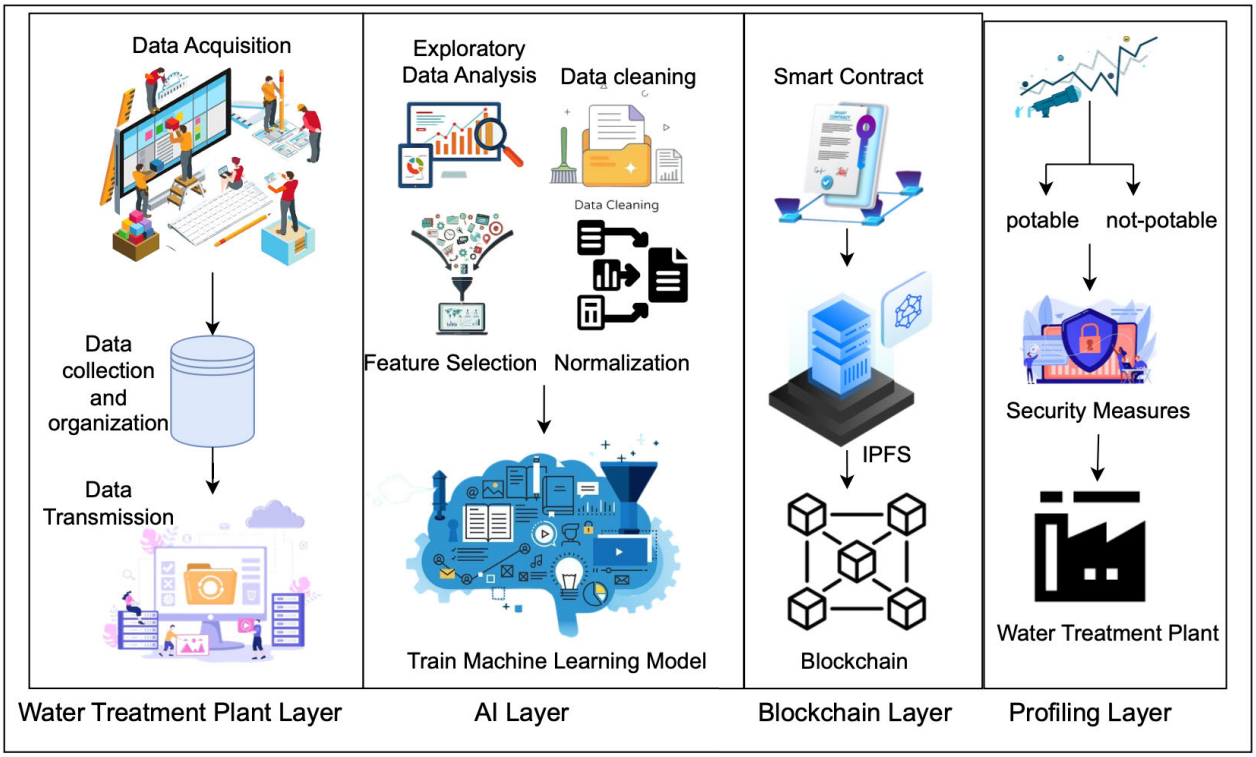


**FIGURE 1.** Proposed model.

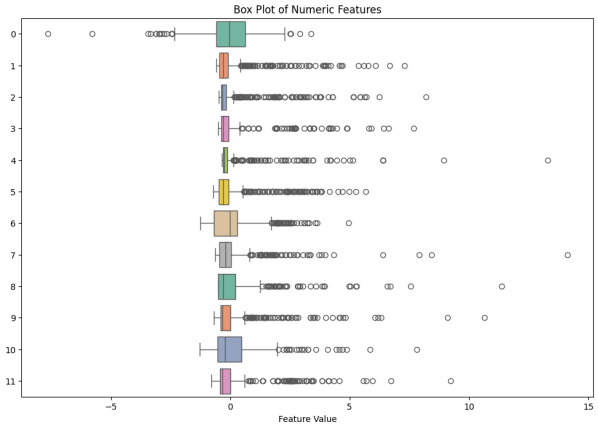


**FIGURE 2.** Box plot of all features.

In the current context, EDA techniques used are statistical analysis, data visualization, and feature engineering. Statistical analysis played a vital role in understanding the characteristics of the collected data and gaining insights into the water parameters that are relevant in water potability prediction. Since we have a classification problem here, we first conducted a chi-square test for independence to determine if there was a significant association between variables. It provided us with chi-squared statistics and corresponding p-values. However, to gain a deeper understanding of the association's magnitude, we also calculated Cramer's V. Cramer's V is a measure of association that

**TABLE 2.** Description of dataset features.

| Feature | Description |
|---|---|
| pH | Level of acidity or alkalinity |
| Sodium | Concentration of sodium ions |
| Magnesium | Concentration of magnesium ions |
| Calcium | Concentration of calcium ions |
| Chloride | Concentration of chloride ions |
| Potassium | Concentration of potassium ions |
| Carbonate | Concentration of carbonate ions |
| Sulphate | Concentration of sulphate ions |
| TDS | Total Dissolved Solids |
| EC | Electrical Conductivity |
| TH | Total Hardness |
| WQI | Water Quality Index |

considers the chi-squared statistic, the total number of observations, and the dimension of the contingency table. The highest value we got by associating each attribute with the potability was the water quality index (WQI), which was very close to 1.

The data visualization techniques were employed to understand the data better visually. Histograms and density plots were used to visualize the range, shape, and spread of the data attributes. Scatter plots were used to explore the relation between numerical relations.

To enhance the performance and predictive power of our model for the water treatment plant, feature engineering was also used. First, relevant information was extracted from the existing features. Numerical features like pH, conductivity and Total dissolved solids(TDS) were scaled and normalized to ensure equal contributions to the model. Missing values

in the dataset were handled through imputation. Feature interactions were explored by creating combinations of features, such as multiplying pH and conductivity, to capture potential interactions affecting water potability. Additionally, domain-specific features like water hardness and alkalinity were calculated based on known relationships. Through these feature engineering steps, our model was able to leverage attributes, which led to improved performance and accurate predictions.

#### 3) MODEL TRAINING

After performing the EDA, the next step involved training the dataset. We took the ratio of the train to test at 80:20, and then we applied the AI algorithms to the train data. The algorithms used are LR, BN, KNN, SGDC, and DT. For each algorithm, the hyperparameters were fine-tuned using grid-search. Grid-search is a systemic method that exhaustively searches through a predefined set of hyper-parameters values to identify the combination that yields the best performance for a given metric, such as accuracy. For example, in LR, some of the hyper-parameters used are regularization parameter(c), penalty, random-stat, and max-iter. Similarly, we used different sets of parameters for different algorithms. This approach helped to maximize the performance of the models by selecting the hyperparameters that yielded the best results for the water potability classification task. FIGURE 3 shows the flow of the AI layer in the proposed model.

Upon evaluating the models, it was observed that DT yielded the highest accuracy, surpassing 0.9944. The algorithm's ability to create hierarchical decision rules based on attribute thresholds proved effective in predicting water potability accurately. Learning curves and ROC curves were plotted to gain a deeper understanding of the performance of the models across the training time frame. They allowed a comprehensive analysis of the model and helped identify the optimal threshold for classification. These evaluations contributed to a robust understanding of the AI models' performance and guided further refinements in the water potability prediction system.

The overall combination of EDA, statistical analysis, feature engineering, and the application of AI algorithms provided an extensive framework for addressing the water treatment plant problem. The findings and insights derived from this process contributed to a strong understanding of water potability prediction and informed decision-making in ensuring that human life expectancy is enhanced.

The DT graph consists of a node in the tree that corresponds to a decision or a feature test. DT make predictions by traversing down the tree, where it evaluates the features at each internal node until it reaches the leaf node. Here, each leaf node decides the final predicted decision class. In our case, the feature responsible for deciding the first split in the decision tree is WQI. This is because it gives the lowest gini-index. Upon plotting the decision tree, a threshold of 49.87 is derived, above which the water is classified as non-potable and below which it is potable for drinking. This approach to decision-making allows for the accurate classification of water samples based on their quality, facilitating informed decisions regarding their potability and suitability for consumption.

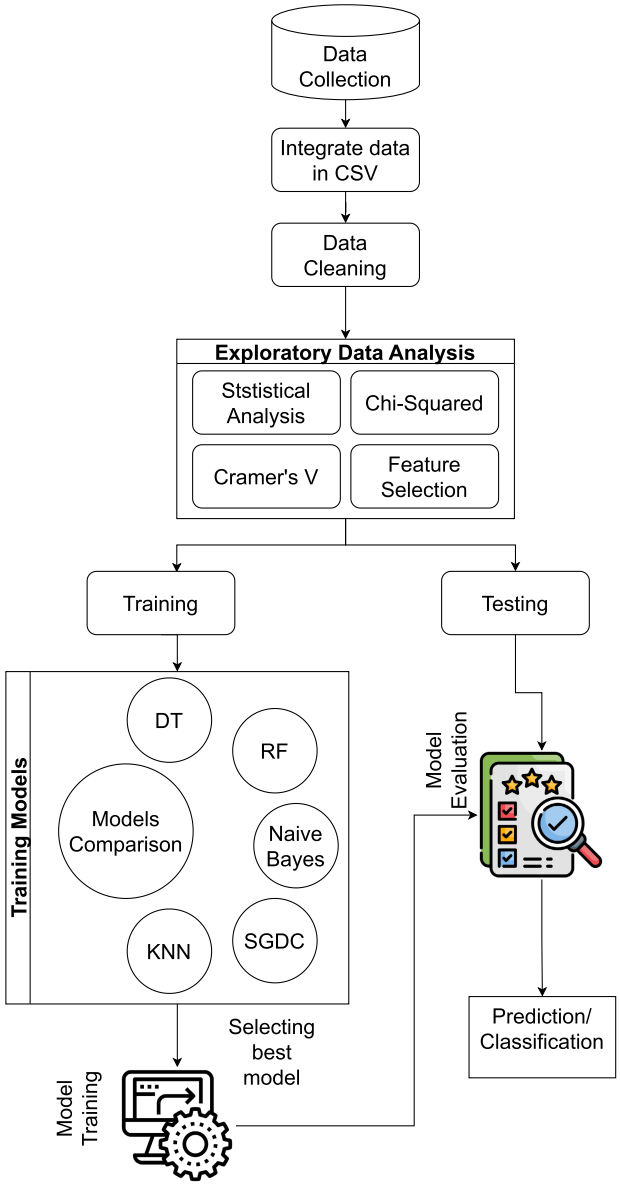


**FIGURE 3.** Working of AI layer.

### C. BLOCKCHAIN LAYER

Blockchain is a technology that consists of records of blocks, and these blocks are used to record transactions across multiple entities of the proposed framework. Blockchain offers a digital, decentralized and immutable ledger that helps to provide security to data stored in blocks. The data from the AI layer is forwarded to this layer, where we implement a digital smart contract that validates the incoming predicted data. The smart contract ensures trust and reliability by eradicating third-party intermediaries from the proposed framework, thus improving the performance of the WTPs. The validated data

**Algorithm 1** Water Treatment Plant Profiling Algorithm

**Data Collection:**
  Collect data from different water treatment plants: $\{w_1, w_2, w_3, \ldots, w_m\} \in W$
  Each plant $w_i$ has sensors $\{s_1, s_2, s_3, \ldots, s_n\} \in S$
  Each sensor $s_i$ collects data $\{\delta_1, \delta_2, \delta_3, \ldots, \delta_n\} \in \Delta$ from $w_i$
**Water Profiling Application:**
  Define a real-time water profiling application $F$
**for all** $w_i \in \{1, 2, \ldots, m\}$ **do**
  **for all** $s_j \in \{1, 2, \ldots, n\}$ **do**
    $w_i(s_j) \xrightarrow[\delta_j]{\text{sends}} F$
  **end for**
**end for**
**Potability Prediction and AI-Based Classification:**
  $DT$ = PredictPotabilityAndClassify($\Delta$)
  Pass the data ($\Delta$) to $DT$ to get the predictions.
  If water is drinkable:
    $F = \delta$
  Else if water is not drinkable:
    $F = \delta'$
**Blockchain Integration:**
  Utilize blockchain for secure data storage.
**Security Measures:**
  Implement security measures to resist attacks.
**Evaluation:**
  Evaluate the proposed model based on performance metrics.
**Results:**
  Proposed model outperforms with accuracy (99.41%).

**Algorithm 2** Predict Potability and Classify

**function** (PredictPotabilityAndClassify Data $\delta$)
  Initialize a Decision Tree classifier $DT$
  Choose the feature that results in the lowest Gini index for the first split
  Set a threshold value based on the chosen feature
  Split the node based on the threshold value
  Recursively repeat the process for each child node until a stopping criterion is met
  Make predictions at the leaf nodes
  **return** $DT$
**end function**

is forwarded to the IPFS (on-site storage), where it computes the hash and is associated with a unique content identifier. Next, the hash is relaid to the blockchain's immutable ledger; this improves the response time of the blockchain network. The rationale behind this is that the raw WTP data is not stored inside the immutable ledger; instead, only the hash is stored. The blockchain members can retrieve the associated data by calling the content identifier from the IPFS web console. Since the size of hash data is relatively lower than the raw data, storing hash data substantially improves the response time of the blockchain in the proposed work.

In the proposed framework, the blockchain layer allows only authorized users, such as water suppliers, the government and entities that use the treated water from WTP, to participate in secured water-based information. It serves as an integrated bridge between government authorities and water suppliers for predicting the effects of polluted water on the human body and life expectancy.

### D. PROFILING LAYER

The data gathered is safely distributed to several WTPs, to improve the water treatment as a whole. Safe sharing guarantees a proper amount of chemical application to non-potable water. Through the utilization of sophisticated profiling methods, the system creates a thorough grasp of characteristics related to water quality. This enables plants to make choices that improve the quality of water by changing water profiling procedures, which in turn helps prolong human life expectancy.

Algorithm 1 describes the entire flow of the proposed blockchain and AI-based model.

## V. RESULTS AND DISCUSSIONS

In this section, we discuss the performance analysis of the proposed framework, which is comprised of AI and blockchain-based results. We considered different performance parameters, such as statistical measures (accuracy, ROC, learning curve) and blockchain's scalability. A detailed explanation of each parameter is as follows.

### A. EXPERIMENTAL SETUP

The proposed framework is implemented in Google Colab's integrated development environment (IDE) with Python 3.10.5. language. To support this language, we used libraries like Numpy, Pandas, Scikit-learn, Seaborn, and Matplotlib. We used pandas for data manipulation and data analysis, which allows us to work with data in a structured way, such as CSV files. Numpy has been used for numerical calculations and to efficiently calculate and evaluate matrices for data analysis and scientific computing. The scikit-learn library is another essential library we use for model selection and evaluation, classification, and data processing. Further, we used the Seaborn library, which is used for data visualization and graph plotting, styling and customizing. The Matplotlib library has been used to plot graphs and curves, which makes it easy for us to visualize data and statistics.

The data coming from the AI layer is forwarded to the Ethereum-based public blockchain, which is implemented in the Remix IDE (0.39.1). It is an online platform for developing and testing smart contracts on the Ethereum-based public blockchain. It helps to write, compile, deploy and debug the smart contracts using the Solidity programming language (0.8.23).

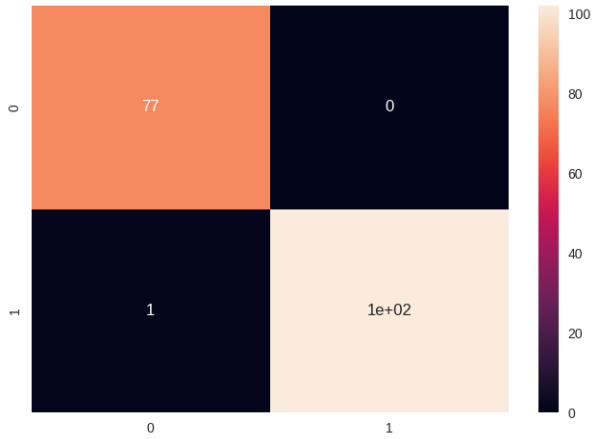


**FIGURE 4.** Confusion matrix.

It helps to create our smart contracts using different functions such as *startPlant*() to start a particular WTP; after starting the plant, only one can add or release the water using *addWater*() and *releaseWater*() functions, once the plant is running *addSample*() and *getSample*() functions are used for profiling of water. These samples contain the details of water parameters, which are important for classifying the water sample into potable and non-potable water. After that, the *stopPlant*() function is used to stop the WTP. Once the smart contract is executed, the authenticated information is transferred to the IPFS-based blockchain ledger. Further, we used calculated precision, recall, and F1-score to calculate the model performance; these are calculated based on the confusion matrix. The learning curve is visualized to understand the model, which shows the rate of learning of the Machine Learning model and the performance of the training and validation datasets. The receiver operating characteristic (ROC) curve is used to evaluate processes in the binary classification model comprehensively.

### B. AI-BASED RESULTS

This section shows different AI-based algorithm and their performance analysis to profile WTP as potable or not potable. To evaluate the performance of the AI algorithms and to understand the results, we need to understand and assess the model based on precision, recall and F1 score. It is important to find the reliability of a model. FIGURE 4 shows the confusion matrix of the DT model. It plots the values of true positive (TP), false positive (FP), true negative (TN) and false Negative (FN). The model classifies the drinkable sample as positive. In contrast to this, negative class classifies the data points that have a negative impact, which in this case are data points that are classified under non-drinkable water. TP measures the extent to which data points are correctly classified under positive. FP measures the extent to which data points are negative and classified as positive. Similarly, FN measures the extent to which negative data points are correctly classified, and TN measures the extent to which negative data are classified as positive. The confusion matrix is used to calculate the precision and the recall.

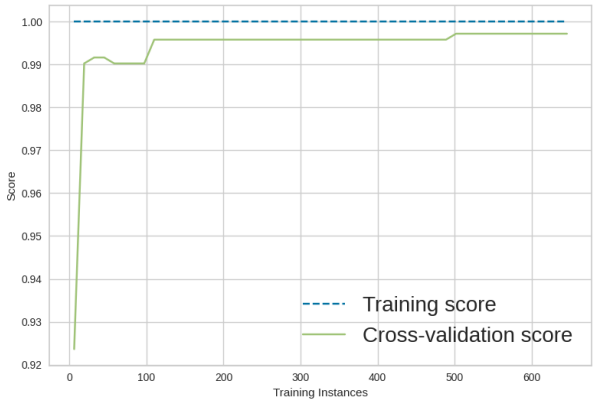


**FIGURE 5.** Learning curve.

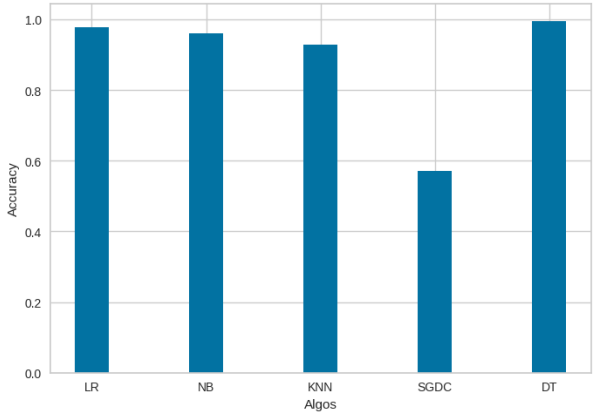


**FIGURE 6.** Accuracy comparison of all machine learning models.

The learning curve shows an analysis of accuracy changes with varying learning efforts. It depicts how a model learns over time graphically; thus, this helps to represent the percentage of rate improvement. This helps to understand the variance of accuracy with the amount of data trained. The steeper the slope, the better the model. From FIGURE 5, we can infer that the model learns incrementally from the data. The accuracy increases with increases in the training data set, giving the final accuracy of 0.9944. FIGURE 6 is a bar graph that compares the accuracy of all the learning algorithms. From the analysis of the curve, we can conclude that DT is getting higher accuracy (99.44%) compared to other algorithms, such as LR, NB, SGDC, and KNN. The reason DT is getting better accuracy is that according to the algorithm, the feature to use as a tree node is a feature that has a very important role in classifying a data point, i.e., WQI.

Precision measures the accuracy of the positive predictions made by the model. It is calculated as the ratio of TP to the sum of TP and FP. Recall measures the ability of the model to identify positive instances correctly. It is calculated as the ratio of TP to the sum of TP and FN predictions. The F1 score has been calculated to find the balance value of both precision and recall. The DT model showcased a high precision value of 1, a recall of 0.990291, and an f1-score of 0.995122.

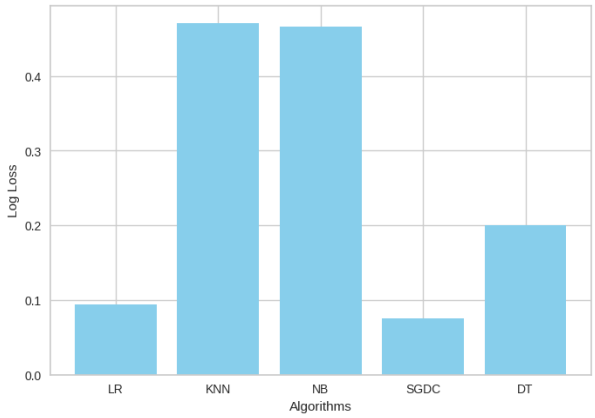


**FIGURE 7.** Log-loss curve comparison.

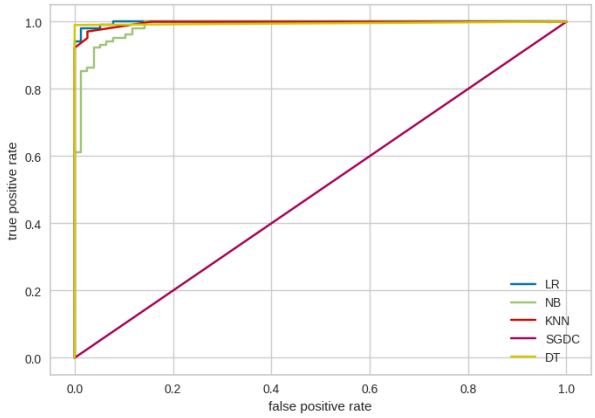


**FIGURE 8.** ROC curve comparison.

A ROC determines the degree of separability of classes. It is a plot between True Positive Rates (TPR) and False Positive Rates (FPR). FIGURE 8 shows the ROC Curve of all the algorithms used in the analysis. The ROC curve was used to represent the probability graph and show the performance of the classification model. The FPR values are plotted on the x-axis, and the TPR values are plotted on the y-axis. A good ROC curve should have high TPR and low FPR, leading to the curve covering the top left corner of the graph. Here, the ROC curve is used to compare the output of different classifiers. By comparing all the curves, we can infer that decision trees have the most accurate results and almost a perfect ROC curve.

The loss evaluation has been done using a log-loss curve. It indicates how close the prediction probability is to the corresponding actual values. FIGURE 7 shows the comparison of log losses of all models. The DT classifier has a log-loss of 0.2, showcasing its exceptional predictive accuracy and reliability. Among the models evaluated, the SGDC classifier demonstrates the lowest log-loss, while the KNN model exhibits the highest log-loss. In addition to evaluating various evaluation metrics of the ML models, it's crucial to consider the training times as they can impact practical implementation. Comparing the training times of the

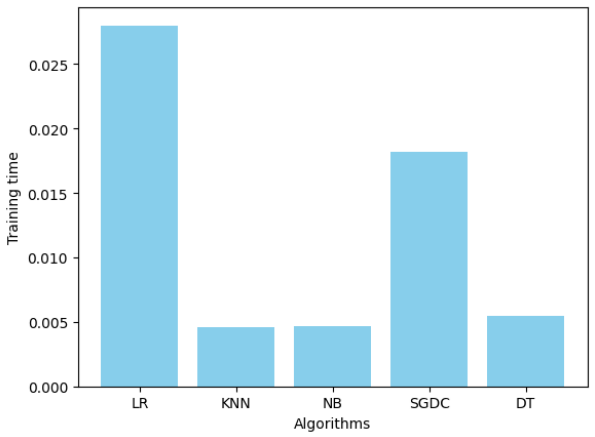


**FIGURE 9.** Training time comparison of ML models.

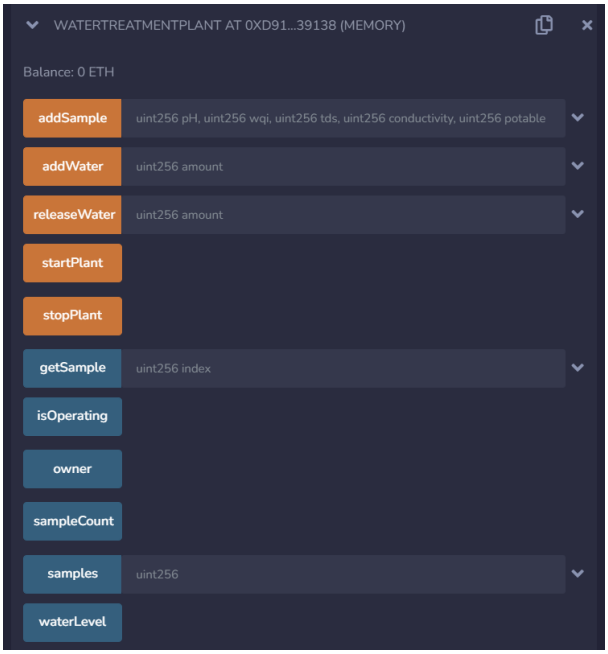


**FIGURE 10.** Smart contract interface.

different algorithms, we find that the algorithms KNN, NB, and DT take comparatively less training time with values of 0.0046, 0.0047, and 0.0054 seconds respectively. As we can see, the model with the best evaluation has a low training time compared to LR and SGDC, which have training times of 0.028 and 0.0182. The comparison is shown in FIGURE 9 using a bar graph.

### C. BLOCKCHAIN-BASED RESULTS

We stored the data in the blockchain network via smart contract. This smart contract is built and deployed using Remix IDE. The interface of the deployed contract is shown in FIGURE 10. This interface contains various functions and variables that are specified in the smart contract such as *startPlant*() to start a particular WTP; after starting the plant only one can add or release the water using *addWater*() and *releaseWater*() functions, once the plant is

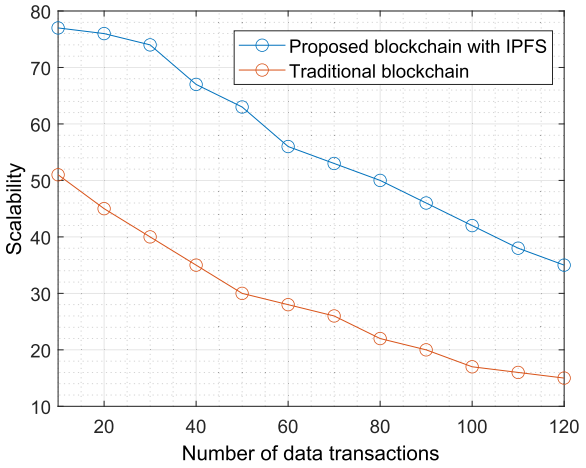


**FIGURE 11.** Scalability comparison with proposed blockchain with IPFS and conventional blockchain.

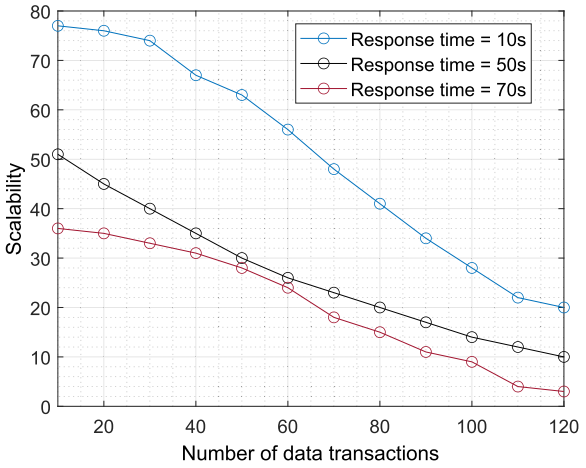


**FIGURE 12.** Scalability comparison with blockchain's response time.

running *addSample*() and *getSample*() functions are used for profiling of water. These samples contain the details of water parameters, which are important for classifying the water sample into potable and non-potable water. After that *stopPlant*() function is used to stop the WTP. It has *isOperating* variable of boolean type to check whether the WTP is in operating mode or not, and whenever we call the *stopPlant*() function, this variable is set to false, which doesn't allow to add or release water functionalities of WTP. Also, this has *waterLevel* variable to check the water level of WTP and whenever we use the functionalities *addWater*() and *releaseWater*(), the value of *waterLevel* will be increased and decreased respectively. *sampleCount* variable is used to check how many samples are being added using *addSample*() functionality. *getSample*() functionality is used to retrieve the sample which is used for profiling. Hence, WTP's operations are controlled using smart contracts.

In the blockchain network, every user runs its node, and nodes communicate with each other for data exchange. The proposed framework utilized IPFS, which computes the hash for the validated data from the smart contract. The hashed data is then forwarded to the blockchain's immutable ledger for an efficient storage and retrieval process. Since raw data is not stored in the ledger, hashed data is stored instead, resulting in a lower processing delay in the blockchain network. As a result, the blockchain node processes more data transactions, improving scalability. FIGURE 11 shows the scalability comparison between the traditional blockchain (not using IPFS) and the proposed blockchain (using IPFS). Based on the aforementioned facts related to the IPFS integration in blockchain, we can infer from the FIGURE 11 that the proposed blockchain has better scalability compared to the traditional blockchain. Additionally, we also compared the blockchain's scalability with different response times (10s, 50s, and 70s) in FIGURE 12. It can be seen from FIGURE 12 that the higher the response time, the lower the scalability, i.e., response time = 70s, scalability = 5 at transaction 120. Conversely, if the response time is lower (due to IPFS), the scalability shows remarkable growth, i.e., response time = 10s, scalability = 20 at transaction 120.

## VI. CONCLUSION

Determining the potability of water presents a significant challenge for researchers due to the influence of external factors. Traditionally, water quality assessment methods rely on subjective analysis and limited attributes. To overcome these limitations, machine learning and deep learning algorithms have gained attention in recent years for the potential of water potability prediction. We have aimed to assess water potability using a machine-learning model. We explored the effectiveness of different machine learning algorithms like DT, KNN, NB, SGDC, and RF. Among the evaluated models, the decision tree algorithms emerged as the most accurate for determining water portability. By leveraging the decision tree algorithm, we can enhance the process of water quality assessment. The reliable classification provided by this model can aid in making informed decisions related to water treatment and resource allocation, ultimately contributing to improved public health and safety. It is important to note that water potability prediction is a complex task influenced by external factors. We plan to advance hybrid models to improve prediction accuracy in future research.

## REFERENCES

[1] Ministry of Jal Shakti. (2022). *Deaths Due to Lack of Clean Water*. Accessed: Mar. 21, 2022. [Online]. Available: https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1807831#:~:text=Composite%20Water%20Management%20Index%20(CWMI,of%20India's%20projected%20population%20by

[2] C. G. Madhusoodhanan, K. G. Sreeja, and T. I. Eldho, "Climate change impact assessments on the water resources of India under extensive human interventions," *Ambio*, vol. 45, no. 6, pp. 725–741, Oct. 2016.

[3] *Water, Sanitation and Hygiene (Wash)*. Accessed: Mar. 14, 2024. [Online]. Available: https://www.who.int/health-topics/water-sanitation-and-hygiene-wash#tab=tab_2

[4] N. Luvhimbi, T. G. Tshitangano, J. T. Mabunda, F. C. Olaniyi, and J. N. Edokpayi, "Water quality assessment and evaluation of human health risk of drinking water from source to point of use at thulamela municipality, limpopo province," *Sci. Rep.*, vol. 12, no. 1, p. 6059, Apr. 2022.

[5] L. Lin, H. Yang, and X. Xu, "Effects of water pollution on human health and disease heterogeneity: A review," *Frontiers Environ. Sci.*, vol. 10, Jun. 2022, Art. no. 880246.

[6] A. H. Stevenson, "Studies of bathing water quality and health," *Amer. J. Public Health Nations Health*, vol. 43, pp. 529–538, May 1953.

[7] M. Marmot, T. Atinmo, T. Byers, J. Chen, T. Hirohata, A. Jackson, W. James, L. Kolonel, S. Kumanyika, and C. Leitzmann, *Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective*. Washington, DC, USA: World Cancer Research Fund/American Institute for Cancer Research, 2007.

[8] T. Page, R. H. Harris, and S. S. Epstein, "Drinking water and cancer mortality in Louisiana," *Science*, vol. 193, no. 4247, pp. 55–57, Jul. 1976.

[9] B. S. Rathi, P. S. Kumar, and D.-V.-N. Vo, "Critical review on hazardous pollutants in water environment: Occurrence, monitoring, fate, removal technologies and risk assessment," *Sci. Total Environ.*, vol. 797, Nov. 2021, Art. no. 149134.

[10] R. Noori, F. Farahani, C. Jun, S. Aradpour, S. M. Bateni, F. Ghazban, M. Hosseinzadeh, M. Maghrebi, M. R. V. Naseh, and S. Abolfathi, "A non-threshold model to estimate carcinogenic risk of nitrate-nitrite in drinking water," *J. Cleaner Prod.*, vol. 363, Aug. 2022, Art. no. 132432.

[11] *Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First and Second Addenda*, World Health Org., Geneva, Switzerland, 2022.

[12] B. Stride, C. Dykes, S. Abolfathi, M. Jimoh, G. D. Bending, and J. Pearson, "Microplastic transport dynamics in surcharging and overflowing manholes," *Sci. Total Environ.*, vol. 899, Nov. 2023, Art. no. 165683.

[13] R. Noori, M. Maghrebi, S. Jessen, S. M. Bateni, E. Heggy, S. Javadi, M. Noury, S. Pistre, S. Abolfathi, and A. AghaKouchak, "Decline in Iran's groundwater recharge," *Nature Commun.*, vol. 14, no. 1, p. 6674, Oct. 2023.

[14] F. Chigondo, B. C. Nyamunda, and V. Bhebhe, "Extraction of water treatment coagulant from locally abundant kaolin clays," *J. Chem.*, vol. 2015, pp. 1–7, Jul. 2015.

[15] Y.-G. Park, W. H. Lee, and K. Kim, "Different adsorption behavior between perfluorohexane sulfonate (PFHxS) and perfluorooctanoic acid (PFOA) on granular activated carbon in full-scale drinking water treatment plants," *Processes*, vol. 9, no. 4, p. 571, Mar. 2021.

[16] P. Guzmán-Avalos, D. Molinero-Hernández, S. Galván-González, N. Herrera-Sandoval, G. Solorio-Díaz, and C. Rubio-Maya, "Numerical design and optimization of a hydraulic micro-turbine adapted to a wastewater treatment plant," *Alexandria Eng. J.*, vol. 62, pp. 555–565, Jan. 2023.

[17] D. Goodarzi, S. Abolfathi, and S. Borzooei, "Modelling solute transport in water disinfection systems: Effects of temperature gradient on the hydraulic and disinfection efficiency of serpentine chlorine contact tanks," *J. Water Process Eng.*, vol. 37, Oct. 2020, Art. no. 101411.

[18] Y. Qiu, S. Ekström, B. Valverde-Pérez, B. F. Smets, J. Climent, C. Domingo-Félez, R. M. Cuenca, and B. G. Plósz, "Numerical modelling of surface aeration and $N_2O$ emission in biological water resource recovery," *Water Res.*, vol. 255, May 2024, Art. no. 121398.

[19] S. J. Sugumar, R. Sahana, S. Phadke, S. Prasad, and G. R. Srilakshmi, "Real time water treatment plant monitoring system using IoT and machine learning approach," in *Proc. Int. Conf. Design Innov. 3Cs Compute Communicate Control (ICDI3C)*, Jun. 2021, pp. 286–289.

[20] F. Granata, S. Papirio, G. Esposito, R. Gargano, and G. De Marinis, "Machine learning algorithms for the forecasting of wastewater quality indicators," *Water*, vol. 9, no. 2, p. 105, Feb. 2017.

[21] U. Safder, J. Kim, G. Pak, G. Rhee, and K. You, "Investigating machine learning applications for effective real-time water quality parameter monitoring in full-scale wastewater treatment plants," *Water*, vol. 14, no. 19, p. 3147, Oct. 2022.

[22] J. Patel, C. Amipara, T. A. Ahanger, K. Ladhva, R. K. Gupta, H. O. Alsaab, Y. S. Althobaiti, and R. Ratna, "A machine learning-based water potability prediction model by using synthetic minority oversampling technique and explainable AI," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–15, Sep. 2022.

[23] A. Alqahtani, M. I. Shah, A. Aldrees, and M. F. Javed, "Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality," *Sustainability*, vol. 14, no. 3, p. 1183, Jan. 2022.

[24] D. Jalal and T. Ezzedine, "Performance analysis of machine learning algorithms for water quality monitoring system," in *Proc. Int. Conf. Internet Things, Embedded Syst. Commun. (IINTEC)*, Dec. 2019, pp. 86–89.

[25] P. Wongburi and J. K. Park, "Prediction of sludge volume index in a wastewater treatment plant using recurrent neural network," *Sustainability*, vol. 14, no. 10, p. 6276, May 2022.

[26] S. Manimekalai, P. B, F. D. Shadrach, V. Lakshmanan, T. Daniya, and T. Guha, "Artificial neural network with extreme learning machine-based wastewater treatment systems," in *Proc. IEEE 2nd Mysore Sub Sect. Int. Conf. (MysuruCon)*, Oct. 2022, pp. 1–6.

[27] K. Banerjee, V. Bali, N. Nawaz, S. Bali, S. Mathur, R. K. Mishra, and S. Rani, "A machine-learning approach for prediction of water contamination using latitude, longitude, and elevation," *Water*, vol. 14, no. 5, p. 728, Feb. 2022.

[28] S. D. Narendar, C. Murugamani, P. R. Kshirsagar, V. Tirth, S. Islam, S. Qaiyum, B. Suneela, M. Al Duhayyim, and Y. A. Waji, "IoT based smart wastewater treatment model for Industry 4.0 using artificial intelligence," *Sci. Program.*, vol. 2022, pp. 1–11, Feb. 2022.

[29] J. Donnelly, A. Daneshkhah, and S. Abolfathi, "Physics-informed neural networks as surrogate models of hydrodynamic simulators," *Sci. Total Environ.*, vol. 912, Feb. 2024, Art. no. 168814.

[30] Y. Gao, W. Xiong, and C. Wang, "Numerical modelling of a dam-regulated river network for balancing water supply and ecological flow downstream," *Water*, vol. 15, no. 10, p. 1962, May 2023.

[31] M. N. Koleva, C. A. Styan, and L. G. Papageorgiou, "Optimisation approaches for the synthesis of water treatment plants," *Comput. Chem. Eng.*, vol. 106, pp. 849–871, Nov. 2017.

[32] K. G. Aparna and R. Swarnalatha, "Dynamic optimization of a wastewater treatment process for sustainable operation using multi-objective genetic algorithm and non-dominated sorting cuckoo search algorithm," *J. Water Process Eng.*, vol. 53, Jul. 2023, Art. no. 103775.

[33] O. Ajayi, A. Bagula, and H. Maluleke, (2022), "Dataset for assessing water quality for drinking and irrigation purposes using machine learning models," *IEEE DataPort*, doi: 10.21227/trcf-1s03.

**DHRUV SARJU THAKKAR** is currently pursuing the Bachelor of Technology degree in computer engineering with the Institute of Technology, Nirma University. His current interests include deep learning, data science, explainable AI, and blockchain.

**ANERI THAKKER** is currently pursuing the Bachelor of Technology degree in computer engineering with the Institute of Technology, Nirma University. Her research interests include deep learning, data science, and blockchain.

**RAJESH GUPTA** (Member, IEEE) received the Bachelor of Engineering degree from the University of Jammu, India, in 2008, the master's degree in technology from Shri Mata Vaishno Devi University, Jammu, India, in 2013, and the Ph.D. degree in computer science and engineering from Nirma University, Ahmedabad, Gujarat, India, in 2023, under the supervision of Dr. Sudeep Tanwar. He is currently an Assistant Professor with Nirma University. He has authored/coauthored some publications (including papers in SCI-indexed journals and IEEE ComSoc-sponsored international conferences). Some of his research findings are published in top-cited journals and conferences, such as IEEE Transactions on Industrial Informatics, IEEE Transactions on Network and Service Management, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Green Communications and Networking, IEEE Transactions on Computational Social Systems, *IEEE Network Magazine*, IEEE Internet of Things Journal, *IEEE Internet of Things Magazine*, *Computer Communications*, *Computer and Electrical Engineering*, *International Journal of Communication Systems* (Wiley), *Transactions on Emerging Telecommunications Technologies* (Wiley), *Physical Communication* (Elsevier), IEEE ICC, IEEE INFOCOM,

IEEE GLOBECOM, IEEE CITS, and many more. His H-index is 31 and i10-index is 67. His research interests include device-to-device communication, network security, blockchain technology, 5G communication networks, and machine learning. He is an Active Member of the ST Research Laboratory (www.sudeeptanwar.in). He was a recipient of the Doctoral Scholarship from the Ministry of Electronics and Information Technology, Government of India, under the Visvesvaraya Ph.D. Scheme. He was a recipient of the Student Travel Grant from WICE-IEEE to attend IEEE ICC 2021 held in Canada. He received the best research paper awards from IEEE ECAI 2021, IEEE ICCCA 2021, IEEE IWCMC 2021, and IEEE SCIoT 2022. His name has been included in the list of Top 2% scientists worldwide published by Stanford University, USA, consecutively in 2021, 2022, and 2023. He attended fully-funded the most prestigious ACM's Heidelberg Laureate Forum 2023 held at Heidelberg University, Germany. He was felicitated by Nirma University for their research achievements bagged, from 2019 to 2022.

**NILESH KUMAR JADAV** (Graduate Student Member, IEEE) received the bachelor's and M.Tech. degrees from Gujarat Technological University (GTU), Gujarat, India, in 2014 and 2018, respectively. He is currently a full-time Ph.D. Research Scholar with the Department of Computer Science and Engineering, Nirma University, Ahmedabad, Gujarat. He has authored/coauthored publications (including papers in SCI-indexed journals and IEEE ComSoc-sponsored international conferences). Some of his research findings are published in top-cited journals and conferences, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, *Digital Communications and Networks* (Elsevier), *Computers and Electrical Engineering* (Elsevier), IEEE INFOCOM, IEEE ICC, and IJCS. His research interests include artificial intelligence, network security, 5G communication networks, and blockchain technology. He is an Active Member of the ST Research Laboratory (www.sudeeptanwar.in).

**SUDEEP TANWAR** (Senior Member, IEEE) is currently a Professor with the Computer Science and Engineering Department, Institute of Technology, Nirma University, India. He is also a Visiting Professor with Jan Wyzykowski University, Polkowice, Poland, and the University of Pitesti, Pitesti, Romania. He has authored two books, edited 13 books, and more than 270 technical papers, including top journals and top conferences, such as IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE WIRELESS COMMUNICATIONS, IEEE NETWORK, ICC, GLOBECOM, and INFOCOM. He initiated the research field of blockchain technology adoption in various verticals, in 2017. His H-index is 68. He actively serves his research communities in various roles. His research interests include blockchain technology, wireless sensor networks, fog computing, smart grids, and the IoT. He is a member of the Technical Committee on Tactile Internet of the IEEE Communication Society. He is a Senior Member of CSI, IAENG, ISTE, and CSTA. He received the Best Research Paper Awards from IEEE GLOBECOM 2018, IEEE ICC 2019, and Springer ICRIC-2019. He has served many international conferences as a member of the organizing committee, such as the Publication Chair for FTNCT-2020, ICCIC 2020, and WiMob2019; a member of the Advisory Board for ICACCT-2021 and ICACI 2020; the Workshop Co-Chair for CIS 2021; and the General Chair for IC4S 2019 and 2020 and ICCSDF 2020. He is serving on the editorial boards for *Frontiers of Blockchain*, *Cyber Security and Applications*, *Computer Communications*, *International Journal of Communication Systems*, and *Security and Privacy*.

**GIOVANNI PAU** (Senior Member, IEEE) received the bachelor's degree in telematic engineering from the University of Catania, Italy, and the master's (cum Laude) and Ph.D. degrees in telematic engineering from the Kore University of Enna, Italy. He is currently an Associate Professor with the Faculty of Engineering and Architecture, Kore University of Enna. He is the author/coauthor of more than 100 refereed papers published in journals and conference proceedings. His research interests include wireless sensor networks, fuzzy logic controllers, intelligent transportation systems, the Internet of Things, smart homes, and network security. He is a member of the IEEE (Italy Section) and has been involved in several international conferences as the session co-chair and a technical program committee member. He serves/served as a leading guest editor for the special issues of several international journals. He is an Editorial Board Member and an Associate Editor of several journals, such as IEEE ACCESS, *Wireless Networks* (Springer), *EURASIP Journal on Wireless Communications and Networking* (Springer), *Wireless Communications and Mobile Computing* (Hindawi), and *Sensors* (MDPI).

**GULSHAN SHARMA** received the B.Tech., M.Tech., and Ph.D. degrees. He is currently a Senior Lecturer with the Department of Electrical Engineering Technology, University of Johannesburg. He is also a Y-Rated Researcher with NRF South Africa. His research interests include power system operation and control and the application of AI techniques to power systems. He is an Academic Editor of *International Transactions on Electrical Energy System* (Wiley) and a Regional Editor of *Recent Advances in Electrical and Electronic Engineering* (Bentham Science).

**FAYEZ ALQAHTANI** is currently a Full Professor with the Software Engineering Department, College of Computer and Information Sciences, King Saud University (KSU). He was appointed as the Director of the Computer Division, Deanship of Student Affairs. He has conducted research projects in several areas of information and communication technology, such as web 2.0, information security, enterprise architecture, software process improvement, the Internet of Things, and fog computing. He has participated in several academic events. He is a member of several academic and professional associations, such as the Association for Computing Machinery (ACM), Australian Computer Society, and the Association for Information Systems.

**AMR TOLBA** (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the Mathematics and Computer Science Department, Faculty of Science, Menoufia University, Egypt, in 2002 and 2006, respectively. He is currently a Full Professor of computer science with King Saud University (KSU), Saudi Arabia. He serves as a technical program committee (TPC) member for several conferences. He served as an associate editor/a guest editor for several ISI journals. He has authored/coauthored over 180 scientific papers in top-ranked (ISI) international journals and conference proceedings. His main research interests include artificial intelligence (AI), the Internet of Things (IoT), data science, and cloud computing.

Open Access funding provided by 'Università degli Studi di Enna "KORE"' within the CRUI CARE Agreement