iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: V Month of publication: May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71046

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Analysis of Water Quality for Potability Using Machine Learning Algorithms(Random Forest)

AVRS Nikhil Kumar[1], Amulya Thipireddy[2], CH. Srinika[3], M. Shoukath Ali[4]
[1, 2, 3]4th, Dept. of ECE, Geethanjali College of Engineering and Technology, India
[4]Associate. Professor, Dept. of ECE, Geethanjali College of Engineering and Technology, India

Abstract: Access to clean and potable water is essential for human health and environmental sustainability. In this study, we present a machine learning-based water quality analysis system designed to determine the potability of water using sensor data and predictive modelling. The system integrates an Arduino Uno microcontroller, interfaced with pH and turbidity sensors, to collect real-time readings from various water samples. This data is then processed and fed into a Random Forest machine learning algorithm, which evaluates multiple parameters—pH level, turbidity concentration, temperature, and algae formation risk—to classify whether the water is safe or unsafe for consumption. The Random Forest model is trained on a dataset of water quality parameters labelled as potable or non-potable, allowing it to make reliable predictions based on incoming sensor readings. By leveraging ensemble learning techniques, the system enhances prediction accuracy, ensuring robust assessment under varying environmental conditions. The real-time nature of data acquisition and analysis makes this system highly effective for continuous monitoring of water quality, providing an automated, scalable solution for water safety evaluation. The findings of this research demonstrate the efficacy of machine learning in environmental monitoring, enabling proactive measures in ensuring safe drinking water. The proposed system can be integrated into industrial water treatment plants, municipal facilities, and remote communities, offering a cost-effective, data-driven approach to water safety management. Future advancements may incorporate cloud-based monitoring, IoT connectivity, and advanced filtration recommendations, further improving water quality assessment frameworks.
Keywords: PH, Turbidity, Arduino, Random Forest.

I. INTRODUCTION

Access to clean and safe drinking water is one of the most essential requirements for human survival. However, a significant portion of the global population faces challenges related to water contamination, which leads to severe health hazards and waterborne diseases. Traditional methods of water quality testing involve laboratory analysis which is often time-consuming, expensive, and impractical for real-time monitoring. The lack of readily available, accessible, and cost-effective solutions for assessing water quality has highlighted the need for a modern, technology-driven approach. This project aims to provide a real-time, automated water quality assessment system that integrates Internet of Things (IoT) technology and machine learning techniques. IoT-based sensors are used to capture critical water quality parameters, such as pH, hardness, and chloramines. These parameters are directly linked to the potability of water, helping determine whether it is safe for human consumption. The collected data is transmitted to a processing unit, converted into a structured CSV format, and analysed using a pre-trained machine learning model. The model predicts the potability of water based on the input parameters, providing a rapid and reliable solution to evaluate water quality. The significance of this project lies in its ability to address the limitations of conventional methods by offering an efficient, scalable, and cost-effective alternative. The IoT-enabled system ensures real- time data collection, while the machine learning model facilitates accurate predictions. Such an integrated approach can be deployed across diverse applications, including households, industries, and communities, enabling proactive monitoring of water quality and minimizing health risks associated with contaminated water. By combining cutting-edge technology with environmental monitoring, this project not only enhances public health safety but also contributes to the larger goal of sustainable water resource management. The use of predictive analytics ensures that potential risks are identified in advance, allowing for timely interventions. Furthermore, the system's modular and scalable design makes it suitable for implementation in various settings, including remote or underserved regions. In essence, this project represents a significant step toward bridging the gap between advanced technology and real-world challenges in water safety.

This project aims to combine Arduino and machine learning to create a simple, low-cost, and intelligent system that can test and analyse water quality, especially useful in areas with limited access to laboratory facilities. Access to clean and safe drinking water is a fundamental human right and a critical public health concern worldwide.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Traditional methods for assessing water quality often rely on laboratory testing, which can be time-consuming, expensive, and inaccessible in remote or under-resourced areas. With the advancement of technology, there is growing interest in the development of low-cost, real-time, and automated systems for water.

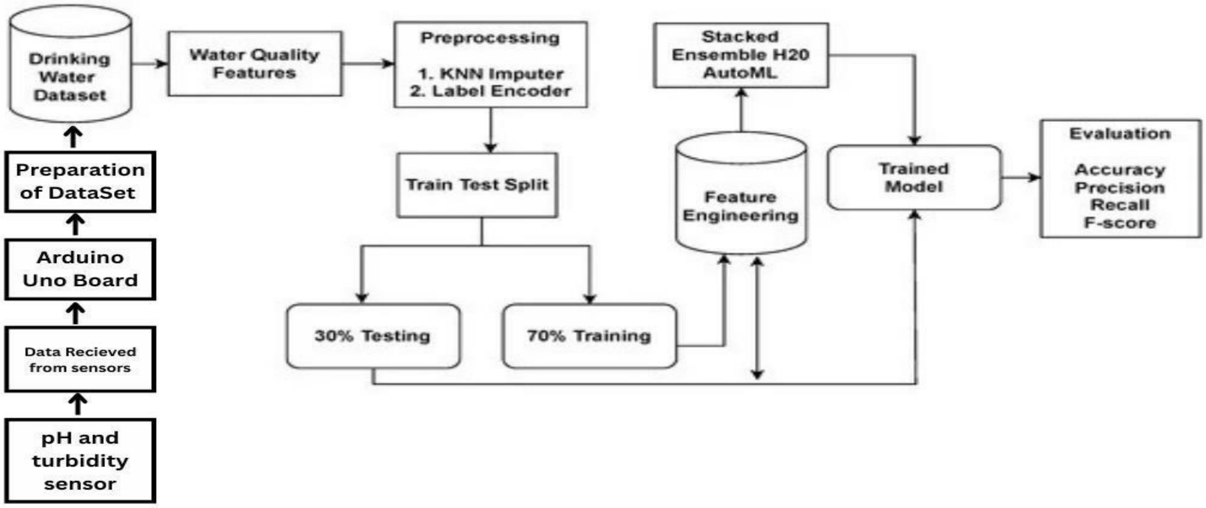


Fig. 1, Block diagram for analysis of water quality using Arduino UNO and Machine learning

II. PROBLEM STATEMENT - PROPOSED SOLUTION

Access to clean and safe drinking water remains a global challenge, with contamination posing serious health risks. Conventional water quality assessment methods rely on manual sampling and laboratory testing, which are often time-consuming, expensive, and impractical for real-time monitoring. Many regions lack readily accessible solutions for evaluating water potability, leading to increased cases of waterborne diseases. The presence of harmful contaminants, excessive turbidity, and improper pH levels can make water unsafe for consumption, necessitating an automated and efficient system for continuous monitoring. Moreover, the early detection of algae formation risks is essential for preventing microbial growth that deteriorates water quality. With increasing concerns over water safety, a data-driven approach to assessing potability is required to make informed decisions about water usage.

To address this issue, we propose an IoT-based water quality monitoring system, leveraging Arduino Uno, sensor technology, and machine learning algorithms for real-time potability assessment. The system integrates pH and turbidity sensors, which collect water sample data, transmitting it to an Arduino microcontroller for processing. This data is then fed into a Random Forest classification model, trained on labelled water quality datasets to predict potability based on multiple factors—pH level, turbidity, temperature, and algae formation risk. The application provides instant feedback, classifying water as either potable or non-potable based on predefined safety criteria. The proposed model utilizes ensemble learning, enhancing prediction accuracy and reliability. The system incorporates a web-based interface, enabling users to visualize real-time sensor readings, historical water trends, and automated alerts for contamination risks. Through wireless connectivity options such as Wi-Fi or Bluetooth, users can remotely track water conditions, making it ideal for households, industries, and water treatment facilities. The implementation of predictive analytics ensures proactive intervention, minimizing the effects of unsafe water consumption and improving public health outcomes. By integrating IoT and AI-driven techniques, this project offers a cost-effective and scalable alternative to traditional water testing methods. It reduces human intervention, enhances response times, and ensures continuous monitoring. Future enhancements may include cloud-based data storage, remote filtration system control, and expanded sensor integration, strengthening the overall efficiency of water quality management. The system's adaptability allows it to be deployed in urban and rural settings, providing communities with a scientific, data-driven approach to water safety assessment.

III. WORKFLOW OF THE PROJECT

The water quality monitoring system is structured to assess the potability of water through an integrated approach involving sensor technology, data transmission, and machine learning analysis. The process begins with data acquisition using Arduino-connected sensors, which measure multiple water quality parameters such as pH, turbidity, conductivity, and temperature. These sensors continuously collect information from water samples, providing real-time insights into potential contamination levels and variations in quality.

Once the sensor data is collected, it is transmitted to a machine learning (ML) model for in-depth analysis. The system employs XGBoost, a powerful ML algorithm known for its accuracy in classification tasks. This model is trained on a dataset consisting of previously recorded water quality values, allowing it to identify trends and predict potability. The ML algorithm uses pattern recognition techniques to determine whether a given water sample meets safety standards based on learned environmental factors. Before feeding the raw data into the ML model, it undergoes preprocessing to ensure accuracy and consistency. This step involves data cleaning, outlier removal, and normalization, optimizing the dataset for analysis. Missing values are handled using imputation techniques, while scaling methods standardize numerical features, ensuring uniformity across different readings. Preprocessing eliminates inconsistencies, thereby improving the model's predictive performance.

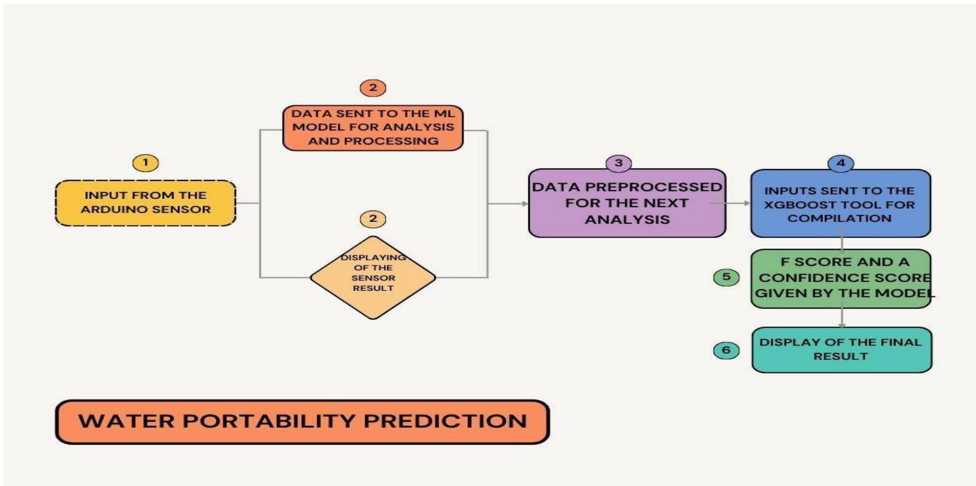


Fig. 2, Workflow steps to implement water quality

The processed data is then sent to XGBoost, where it is compiled and used for water quality classification. XGBoost operates through gradient boosting, refining predictions by integrating multiple decision trees. This results in highly accurate assessments, making it an ideal choice for determining whether a given water sample is safe or unsafe for consumption. The model evaluates factors such as pH stability, turbidity levels, and temperature fluctuations, which collectively impact water potability. To ensure model reliability, performance evaluation metrics such as the F-score and confidence score are generated. The F-score measures the balance between precision and recall, ensuring minimal false positives and negatives in water safety classification. The confidence score, on the other hand, reflects how certain the model is about its prediction. High confidence values indicate strong certainty, while lower scores may suggest borderline classifications requiring further verification. Finally, the system displays the classification results, providing users with clear indications of water quality. The output may be a straightforward "safe" or "unsafe" determination, or it may include additional details regarding contaminants, turbidity fluctuations, and chemical stability. By integrating real-time sensor data, predictive analytics, and interactive results visualization, this system offers a data-driven approach to water quality assessment, ensuring proactive monitoring for improved safety.

IV. WATER QUALITY HARDWARE DESIGN USING ARDUINO UNO

The hardware implementation of the water quality monitoring system involves an Arduino Uno microcontroller interfaced with pH and turbidity sensors to collect real-time water quality data. The Arduino Uno, known for its versatility and ease of integration, serves as the central processing unit, handling data acquisition and transmission. This microcontroller is equipped with digital and analog input pins, enabling seamless communication with multiple sensors for efficient data processing. The system is designed to continuously monitor water samples and determine potability based on predefined thresholds. The pH sensor measures the acidity or alkalinity of the water, providing readings on a scale of 0 to 14, where values below 7 indicate acidity, and values above 7 indicate alkalinity. Proper pH levels are critical in water safety, as extreme variations may indicate contamination or the presence of harmful substances. The turbidity sensor, on the other hand, assesses the cloudiness or haziness of the water by measuring the light scattering effect caused by suspended particles. Higher turbidity values suggest increased contamination, requiring filtration or treatment before consumption. These sensors transmit their readings as analog signals, which are processed by the Arduino using its built-in analog-to-digital converter (ADC). To ensure efficient data handling, the Arduino Uno is programmed using the Arduino IDE, incorporating necessary calibration functions to improve accuracy.

Sensor readings are collected at regular intervals and transmitted via serial communication to a connected computer or wireless modules such as ESP8266 or Bluetooth modules, enabling real-time monitoring. The data transmission system allows for wired connections using USB interfaces or remote logging through Wi-Fi, facilitating extended deployment for continuous water quality assessment. A power management system is integrated to ensure stable operation, using either DC power supply or battery backup for uninterrupted functionality. The sensors are powered through the Arduino's 5V and 3.3V output pins, preventing voltage fluctuations that could affect data accuracy. Additionally, an LCD display module may be connected to the setup to provide instant on-site feedback, allowing users to view live water quality readings without requiring external software interfaces. The completed hardware setup forms a comprehensive water quality monitoring system, leveraging sensor integration, microcontroller processing, and wireless data transmission to ensure reliable potability assessment. The system is designed to be cost-effective, scalable, and adaptable for diverse applications, including household, industrial, agricultural, and municipal water treatment facilities. Future enhancements may involve cloud-based monitoring, automated filtration control, and AI-driven contamination detection, further strengthening water safety measures for sustainable environmental health.

V. INTEGRATING MACHINE LEARNING WITH HARDWARE

Integrating machine learning with hardware for water quality analysis is a transformative approach that enhances real-time monitoring and predictive capabilities. This system utilizes Arduino Uno, which interfaces with pH and turbidity sensors to collect raw water quality data. The Arduino is connected to a computer via a USB cable, which serves both as a power source and a communication bridge for data transmission. Once the sensors are submerged in different water samples, they detect variations in pH levels and turbidity, which are crucial indicators of potability. The readings are sent to a Python-based machine learning model, leveraging Random Forest classification to determine whether the water is safe or unsafe for consumption. The machine learning integration is facilitated using Streamlit, a Python framework that provides an interactive user interface for visualizing real-time sensor data. The collected water quality parameters, including pH, turbidity, temperature, and algae formation risk, undergo preprocessing before being fed into the Random Forest model. This model is trained on labeled datasets to analyze relationships between water composition and potability, ensuring precise classification. The predictions are displayed on a local web interface, allowing users to view whether the water sample meets potability standards or requires treatment. The Random Forest algorithm plays a critical role in this system by employing an ensemble learning approach, which combines multiple decision trees to enhance predictive accuracy. By analyzing sensor data trends, it can detect minor fluctuations in water composition that may indicate contamination risks. Additionally, the model computes an F-score and confidence score, ensuring reliable water quality assessments. The results are categorized as "Safe to Drink" or "Not Potable", accompanied by explanations highlighting deviations in pH levels and turbidity thresholds. Using real-time data streaming, the system provides continuous monitoring with automated updates every second. This is achieved through the auto-refresh mechanism in Streamlit, which ensures that new sensor readings are instantly processed and displayed. By integrating IoT connectivity, the model can extend its capabilities to cloud-based monitoring and remote tracking, allowing users to access historical trends and take corrective measures for improving water purity. Overall, this hardware-software integration represents a cost-effective, scalable solution for water quality assessment. It eliminates the need for manual testing, reduces response times, and enhances data-driven decision-making. Future advancements could incorporate additional sensors for broader contamination detection, AI-driven filtration recommendations, and predictive maintenance algorithms, further optimizing water safety protocols. This project provides a scientific and automated approach to safeguard public health by ensuring access to clean drinking water.

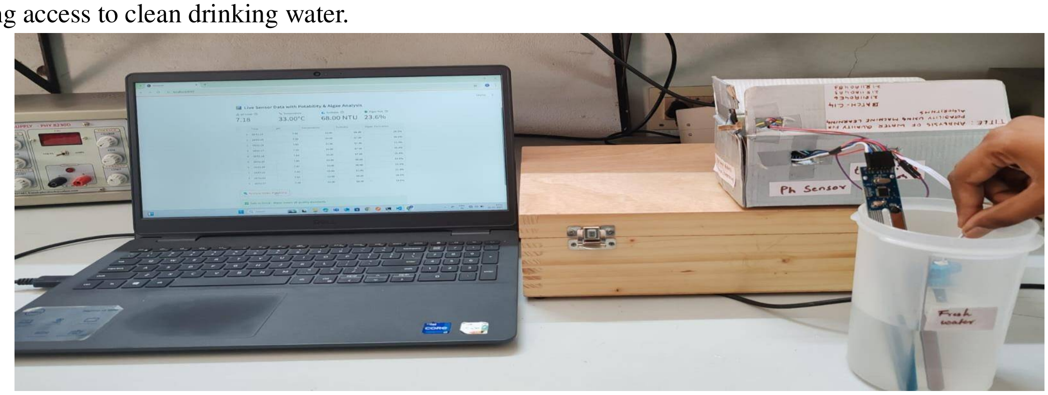


Fig. 4, The final product to check water portability

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

VI. RESULT AND DISCUSSION

We conducted water quality tests on three distinct water samples—freshwater, lemon water, and muddy water—using pH and turbidity sensors integrated with the Arduino Uno. Each sample was analysed to determine its pH level and turbidity, which were then processed through a machine learning model based on Random Forest classification to assess its potability. Below are the detailed findings for each case.

Fresh Water Sample:

pH Level	Temperature	Turbidity	Algae Risk
7.00	33.00°C	69.00 NTU	25.6%

	Time	pH	Temperature	Turbidity	Algae_Formation
0	16:54:26	7.00	33.00	69.00	25.6%
1	16:54:27	7.00	33.00	69.00	25.6%
2	16:54:28	7.02	33.00	69.00	25.4%
3	16:54:29	7.00	33.00	69.00	25.6%
4	16:54:39	6.97	33.00	69.00	26.0%
5	16:54:40	7.00	33.00	69.00	25.6%
6	16:54:41	7.00	33.00	69.00	25.6%

Fig 5, Analysis of Water Potability using Fresh Water Sample

The freshwater sample recorded a pH range of 5.5 to 7, which falls within the acceptable range for potable water. A pH closer to 7 is considered neutral and indicates that the water is free from excessive acidity or alkalinity, making it safe for consumption. The turbidity level was also found to be low, suggesting minimal suspended particles and adequate clarity. Based on these observations, the machine learning model classified the sample as potable, confirming that freshwater is safe for drinking. The sensor readings aligned with expected values, reinforcing the accuracy of the system in determining water quality parameters.

A. Lemon Water Sample

pH Level	Temperature	Turbidity	Algae Risk
16.03	35.00°C	43.00 NTU	33.2%

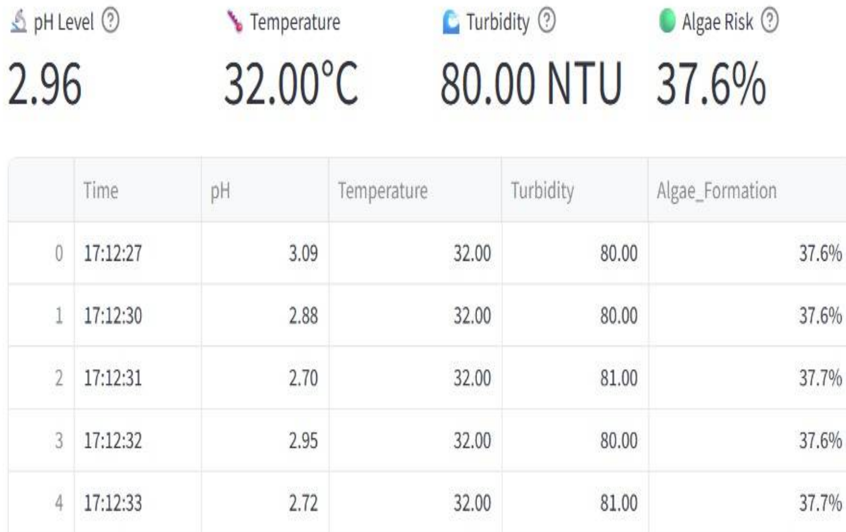
	Time	pH	Temperature	Turbidity	Algae_Formation
0	16:52:21	6.97	34.00	49.00	23.6%
1	16:52:22	6.97	33.00	67.00	25.7%
2	16:52:23	16.03	35.00	43.00	33.2%

Fig 6, Analysis of Water Potability using Lemon Water Sample

The lemon water sample exhibited a pH level greater than 7, indicating that the solution was highly acidic due to the presence of citric acid. Acidic water conditions can lead to corrosive effects on plumbing systems, impacting water distribution infrastructure. While naturally occurring acidity in food items such as lemon juice is acceptable, water meant for drinking should generally fall within the range of 6.5 to 8.5 for safety. The turbidity levels in lemon water remained low, as there were no excessive suspended particles. However, due to its high acidity, the system classified the sample as non-potable, recommending further neutralization or treatment before consumption.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

B. Muddy Water Sample



pH Level 2.96 | Temperature 32.00°C | Turbidity 80.00 NTU | Algae Risk 37.6%

	Time	pH	Temperature	Turbidity	Algae_Formation
0	17:12:27	3.09	32.00	80.00	37.6%
1	17:12:30	2.88	32.00	80.00	37.6%
2	17:12:31	2.70	32.00	81.00	37.7%
3	17:12:32	2.95	32.00	80.00	37.6%
4	17:12:33	2.72	32.00	81.00	37.7%

Fig 7, Analysis of Water Potability using Dirty Water Sample

The muddy water sample presented a pH value below 3, reflecting extreme acidity that poses significant health risks if consumed. Additionally, the turbidity levels were exceptionally high, indicating dense suspended particles and impurities, which further degraded water quality. High turbidity levels suggest the presence of organic matter, sediment, and contaminants, making the water unsafe for direct consumption. Based on these readings, the machine learning model classified muddy water as non-potable, requiring filtration, chemical treatment, or purification before use. These findings validated the reliability of sensor-based assessments, confirming that highly contaminated water should undergo thorough processing and purification before being deemed suitable for drinking.

The results from this water quality detection using hardware and machine learning highlight the effectiveness of Arduino-based water quality monitoring in assessing water potability using sensor data and machine learning models. Freshwater, with its pH range between 5.5 and 7, was classified as safe for drinking, while lemon water and muddy water failed the potability test due to high acidity and excessive turbidity, respectively. This approach provides real-time, automated classification, enabling proactive water quality assessments for households, industries, and municipal water treatment plants. Future enhancements, such as cloud-based monitoring and automated filtration recommendations, can further improve water quality analysis and safety protocols.

VII. CONCLUSION

The successful implementation of the Water Potability Prediction system highlights the effectiveness of combining real-time sensor data, Arduino-based hardware, and machine learning algorithms for assessing water quality. By integrating pH and turbidity sensors with an Arduino Uno microcontroller, the system captures essential water parameters, which are then processed using a Random Forest classifier. This classification model efficiently determines whether water is potable or non-potable, offering an automated solution for continuous monitoring. The real-time predictions reduce human intervention and error, ensuring that users receive instant feedback about the safety of their drinking water. This approach is particularly beneficial in resource-limited regions, where manual water testing is impractical or infrequent. While the system successfully classifies water potability, some limitations remain. One key limitation is the restricted number of sensor parameters being monitored. Expanding the system to incorporate additional sensors—such as temperature, dissolved oxygen, and conductivity meters—could provide a more detailed and comprehensive assessment of water quality. Another challenge lies in the accuracy of machine learning predictions, which heavily depend on the quality and diversity of training datasets. If the dataset lacks wide-ranging water samples or environmental variations, the model's predictions may be less reliable, potentially affecting potability classification. Looking forward, there is substantial potential for system enhancements through advanced technologies. Cloud integration would allow for remote data access, centralized storage, and long-term analytics, enabling authorities and consumers to track water quality trends over time. Additionally, incorporating edge computing techniques could process sensor data locally, reducing latency in predictions and enabling instant decision-making.

By leveraging AI-powered filtration recommendations, the system could provide adaptive solutions, suggesting appropriate water treatment measures based on contamination levels. Further improvements could also include the development of mobile applications, making real-time potability assessments more accessible. A user-friendly app could enable individuals to receive instant alerts and notifications, informing them about potential contamination risks or required corrective actions. This feature would be particularly useful for households, industries, and municipal water providers, ensuring proactive water management without dependence on manual testing. The Water Potability Prediction system presents a scalable and cost-effective framework for automated water safety assessment. With continued refinements in sensor integration, AI-based classification, and cloud-enabled monitoring, this approach can revolutionize water quality management, providing a smart and reliable solution for safe drinking water worldwide. Let me know if you need adjustments or a formatted journal-ready version

REFERENCES

[1] Hussein, E. A., Thron, C., Ghaziasgar, M., Bagula, A., & Vaccari, M. (2020). Groundwater prediction using machine-learning tools. Algorithms, 13, 1–16. https://doi.org/10.3390/a13110300

[2] Jha, D. K., Devi, M. P., Vidyalakshmi, R., Brindha, B., Vinithkumar, N. V., & Kirubagaran, R. (2015). Water quality assessment using water quality index and geographical information system methods in the coastal waters of Andaman Sea, India. Marine Pollution Bulletin, 100, 555–561. https://doi.org/10.1016/j. marpolbul.2015.08.032

[3] Kalini, Z., International Journal of Information Management A SEM-neural network approach for predicting antecedents of m-commerce acceptance, 37 (2017) 14–24. https://doi.org/10.1016/j.ijinfomgt.2016.10.008.

[4] Kar, A. K., & Varsha, P. S. (2023). A review and research agenda. International Journal of Information Management Data Insights. , Article 100176. https://doi.org/10.1016/j. jjimei.2023.100176

[5] Kothari, N., Shreemali, J., Chakrabarti, P., & Poddar, S. (2021). Design and implementation of IoT sensor based drinking water quality measurement system. Materials Today: Proceedings. https://doi.org/10.1016/j.matpr.2020.12.1142

[6] Lamare, R.E., Singh, O.P., Application of Ccme water quality index in evaluating the water quality status in limestone mining area of, 10 (2016) 149–154. Liu, P., Wang, J., Sangaiah, A., Xie, Y., & Yin, X. (2019). Analysis and prediction of water quality using LL.S.T.M.Deep neural networks in IoT environment. Sustainability, 11, 2058. https://doi.org/10.3390/su11072058

[7] Lumb, A., Sharma, T. C., & Bibeault, J.-F. (2011). A review of genesis and evolution of water quality index (WWQI and some future directions. Water Quality, Exposure and Health: Purpose and Goals, 3, 11–24. https://doi.org/10.1007/s12403-011-0040-0

[8] Mancuso, P., Piccialli, V., & Sudoso, A. M. (2021). A machine learning approach for forecasting hierarchical time series. Expert Systems with Applications, 182, Article 115102. https://doi.org/10.1016/j.eswa.2021.115102

[9] Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578. https://doi. org/10.1016/j.jhydrol.2019.124084

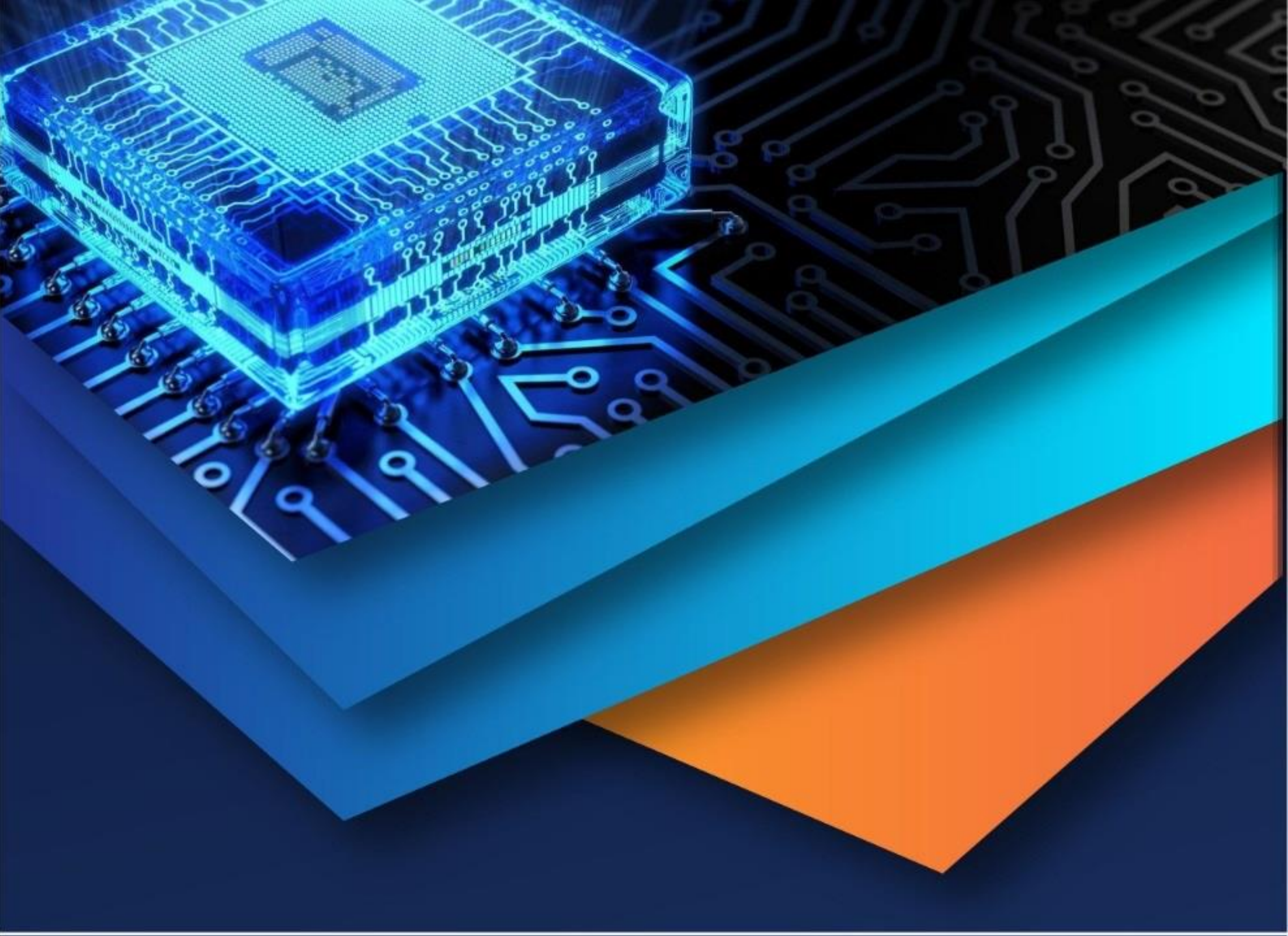
[10] Nampak, H., Pradhan, B., & Manap, M. A. (2014). Application of GGISbased data driven evidential belief function model to predict groundwater potential zonation. Journal of Hydrology, 513, 283–300. https://doi.org/10.1016/j.jhydrol.2014.02.053

[11] Nayak, J. G., Patil, L. G., & Patki, V. K. (2020). Development of water quality index for Godavari River (India) based on fuzzy inference system. Groundwater for Sustainable Development, 10, Article 100350. https://doi.org/10.1016/j.gsd.2020.100350

[12] Nayar, R., Dr. (2020). Assessment of water quality index and monitoring of pollutants by physico-chemical analysis in water bodies: A review. International Journal of Engineering Research, 9, 178–185. https://doi.org/10.17577/ijertv9is010046

[13] Pasika, S., & Gandla, S. T. (2020). Smart water quality monitoring system with cost- effective using IoT. Heliyon, 6, e04096. https://doi.org/10.1016/j.heliyon.2020. e04096

[14] Roseela, J. A., Godhavari, T., Narayanan, R. M., & Madhuri, P. L. (2020). Design and deployment of IoT based underwater wireless communication system using electronic sensors and materials. Materials Today: Proceedings, 45, 6229–6233. https://doi.org/10.1016/j.matpr.2020.10.586



10.22214/IJRASET

INDEX COPERNICUS 45.98

ISRA JIF IMPACT FACTOR: 7.129

SJIFactor TOGETHER WE REACH THE GOAL IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY