IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Smart Water Quality Prediction System Using Sensor Data And Machine Learning Regression Techniques

P Akhilesh, D Naveen Kumar, Ediga Lokesh Goud, T G Hemanth, Dr. S. Revathy, Dr. R. Shoba Rani, Dr. F. Jerald

Students: Department of Artificial Intelligence
Assistant Professor: Department of Artificial Intelligence
Professor: Department of Artificial Intelligence
Professor: Department of Artificial Intelligence

DR M G R Educational and Research Institute, Chennai 600095, India

ABSTRACT

The stream of water needs to be monitored to ensure the sustainability of the environment, industrial safety, and the overall health of the population. The traditional methods of monitoring rely on a chemical analysis of samples in the laboratory which is expensive, time-consuming and cannot be used in real time. The development of Internet of Things technology (IoT) and machine learning algorithms enables intelligent systems to forecast the indexes of water quality effectively with continuous sensor streams of information.

The paper provides a proposal of Smart Water Quality Prediction System, which can combine real-time sensor measurements and machine learning regression to estimate water quality parameters and Water Quality Index (WQI) in general. The parameters that are processed by the system include turbidity, pH, electrical conductivity, temperature and dissolved oxygen (DO). Various regression algorithms are done and compared based on their performance measures like MAE, MSE, RMSE, and R^2 score depending on the type of regression algorithm used like Linear Regression, Support Vector Regression (SVR) and Random Forest Regression. The experimental findings show that the Random Forest Regression model has a higher accuracy in its prediction whereby the R^2 value is 0.96 with a low error in making predictions. The system that is proposed presents a scalable and smart solution to real time environmental monitoring and decision support system at a low cost.

Keywords:-Machine Learning, Water Quality Index, Water Quality Prediction, Historical Sensor Data, Regression Analysis, Random Forest.

I. INTRODUCTION

The high rate of industrialization, agricultural run-off, urban wastewater discharge as well as poor disposal of wastes have become a major global problem since water pollution. Polluted water has an impact on the ecosystems, drinking water and human health. Surveillance of water quality in rivers, lakes, reservoirs and distribution systems is then of critical importance.

Conventional water quality evaluation involves sampling and laboratory analyses every now and then. These methods are correct but they do not have capability of real-time and predictive intelligence. Also, lab tests are expensive to run and time consuming in favor of analysis.

Recent developments in machine learning allow assessing environmental data automatically. Regression analysis is also particularly relevant to the prediction of continuous water quality parameters. These models can be used with IoT sensors to provide intelligent and continuous monitoring systems.

The present study hypothesizes a Smart Water Quality Prediction System with end-to-end, which is able to:

- Online collection of data on IoT sensors.
- Smart regression-based forecasting.
- Comparison of several ML models.
- Smart environmental monitoring scalable deployment.

II. RELATED WORK

Because pollution keeps growing, watching over water quality matters more now. What began as simple machine learning tied to IoT devices slowly shaped how we track changes. Ways of checking water health evolved when smart sensors started feeding data into early prediction tools. Growing demands for clean resources pushed these methods forward without grand plans. Over time, small tech steps added up to stronger oversight across whole systems.

Out in the field, sensors track water conditions while linear regression makes sense of the numbers on the fly. Though straightforward and cheap, that combo stumbles when nature doesn't follow a straight line [1]. Tossing several smart algorithms together instead - hybrid methods - handles messy patterns better. Accuracy climbs because each model covers gaps left by others, making results steadier overall [2].

Running on remote servers, these water forecasts use grouped math tricks that handle big data faster because number-crunching spreads across many machines at once [3]. Heavy metals hiding in water get spotted just as well when smart algorithms learn odd clues regular tests miss [4].

Simple models like Linear Regression or SVR get picked a lot - they're fast, easy to run, but often fall short when patterns twist too much [5]. When curves get wild, ensemble methods step in - mixing predictions from several models helps dodge overfitting while lifting accuracy [6]. Looking back at past studies, these smart blends stand out especially in tracking messy real-world signals, pulling useful links from big piles of environmental numbers [7].

Most regression tools used to forecast river quality depend heavily on how well the data is cleaned and which variables are picked - this shapes how accurate results turn out [8]. Looking at past patterns over time helps guess what comes next, though these methods can stumble when nature shifts fast or behaves unpredictably [9].

Fitted with machine learning, IoT monitoring tools gather live data - offering steady forecasts that boost performance yet raise expenses and intricacy [10].

Because older methods struggle, deep learning steps in to help. What makes hybrid setups like CNN-LSTM work well is their ability to track patterns across space and time at once. Better predictions come out of that shift - accuracy climbs as a result [11]. When it comes to tuning settings, smarter adjustments lift results further, especially when forecasting dissolved oxygen levels [12].

Last time we checked, LSTM networks handle sequences well. Because they track connections across steps, predictions get sharper [13]. Some models pair them with CNNs - this mix pulls out fine details while still watching how things shift over time. Combining these layers lifts results higher, simply by doing two jobs at once [14].

One way to boost performance is using attention inside deep networks, making it easier for models to spotlight key patterns across time. Instead of treating all data points equally, methods like Bi-LSTM paired with attention weigh inputs by importance - this helps when tracking messy, changing water conditions [15].

Even with progress, problems stick around - like steep computing demands, reliance on massive data, hurdles in scaling - all pointing toward simpler, cheaper ways to predict things.

III. PROBLEM DEFINITION AND OBJECTIVES OF THE RESEARCH

The water quality course of action provides an estimate of a target variable most often a continuous one WQI or dissolved oxygen level as a result of a set of variables that have been determined to known to depend on the target.

Key challenges include:

- Parameters of water No linear relationships.
- Sensor noise, missing values.
- Correlation in the characteristics of inputs.
- Over fitting of regression model.

Research objectives:

- Develop a relevant generated prediction system of regression.
- Compare multiple regression algorithm results.
- Tune the Hyperparameters so that they can be optimized to improve performance.
- confirm scalability of systems in order to release the same in real time operation.

IV. PROPOSED METHODOLOGY

It is represented in the methodology presented below that is broken down into five main stages that are detailed in the following manner:

1. Data Acquisition.
2. Cleanup and Preprocessing of Data.
3. Feature Engineering.
4. Model Training.
5. Performance Evaluation.

A. Data Acquisition

IoT sensors continuously measure:

- pH (acidity/alkalinity level)
- Turbidity (water clarity)
- Temperature
- Dissolved Oxygen (DO)
- Electrical Conductivity (EC)

These sensors transmit the data in the form of the central processor unit as the receiver with the help of the wireless communication protocols.

B. Data Preprocessing

In preprocessing of data, it helps in ensuring that the input of the model is high quality.
Steps include:

- Missing value imputation
- Z-score normalization
- Elimination of outliers by the use of IQR.
- Min Max scaling (Feature scaling)

Normalization formula:
[
X' = \frac{X - X{min}}{X{max} - X{min}}
]

C. Regression Models

1. Linear Regression

One of the simplest machine learning algorithms is linear Regression which is employed in prediction. It develops a linear relationship between the input variables e.g. pH, temperature and turbidity and the output variable. The model is used to predict values with the use of a best-fit straight line which minimized the error between the actual and the predicted values. This is due to the fact that it is a simple model that is used as a baseline given that it is not capable of capturing complicated patterns in the water quality data.

- **Pros:** Easy to apply; Cheap to compute.
- **Limitations:** Nonlinear data not able to be processed, Outliers are sensitive.

2. Support Vector Regression (SVR)

Support Vector Regression is a regression methodology that relies on the Support Vector Machines and can fit data within a given error margin. It is based on the model to represent both linear and nonlinear associations by using kernel functions. It is proper to use it in small to medium data and is able to retrieve complicated patterns but needs a delicate parameter tuning.

- **Pros:** Nonlinear data; Is good in generalization.
- **Limitations:** There is a need in the tuning of the parameters; Slower in large data.

3. Random Forest Regression

Random Forest Regression is also an ensemble learning, which involves using a number of decision trees to enhance the accuracy of the prediction. Every tree is trained with varying data subsets and average results are obtained to get the ultimate result. It is highly adequate in dealing with nonlinear relationships and is more precise when dealing with complex data.

- **Pros**: Is very accurate; The nonlinear relationships.
- **Limitations:** No faster; Not as much understandable.

V. SYSTEM ARCHITECTURE

The architectural building involves:

- Sensor Layer
- Data Processing Layer
- Machine Learning Engine
- Visualization Dashboard
- Alert System

1. Sensor Layer

The sensor layer will be in charge of measuring the parameters of water quality including pH, dissolved oxygen (DO) and turbidity. These values in this system are provided by the IoT sensors or simulated by past data. The resulting data is then transmitted into the succeeding layer.

2. Data Processing Layer

This layer is used to train, serve, and interpret raw data. It entails cleaning of data, dealing with missing data and elimination of inconsistencies. Normalization and feature extraction is implemented so that a data is uniform, and noisy, and it can be used to train the model.

3. Machine Learning Engine

The main component of the system is the machine learning engine. It employs regression algorithms like Linear Regression, Support Vector Regression (SVR) and Random Forest Regression. These models are run based on the history information to predict water quality parameters and WQI.

4. Visualization Dashboard

The depiction dashboard shows the results prediction as graphs, charts, and numerical values. It makes it easy to know the water quality status and make sound decisions by users, researchers and authorities.

5. Alert System

This alarm system tracks the estimated values and sounds out warnings when they surmount the safe targets. This allows a prompt reaction and prevents the possible health and environmental threats.

➢ Overall Workflow

The system is a sensor data collection system which then processes the sensor data and passes it into machine learning models to predict. The index of predicted water quality is seen on the dashboard, and warnings are issued in case unsafe conditions are detected. The real-time monitoring and decision making are made efficient in this structure.

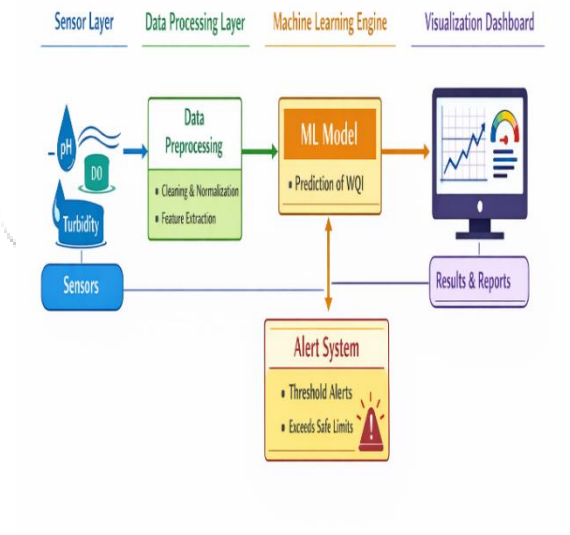

Fig No: 1 - SYSTEM ARCHITECTURE

Flow Diagram:

Sensors - Data Preprocessing - Machine Learning (ML) Model - Predict WQI- Alert System.

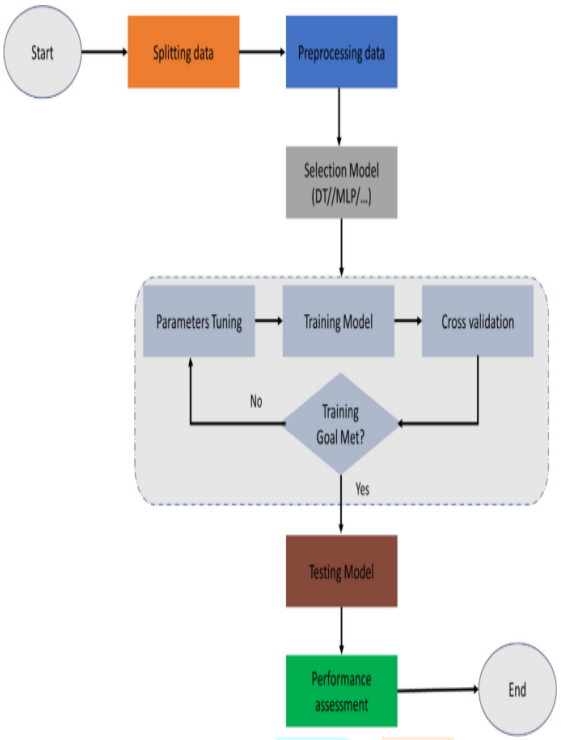

Fig No: 2 -FLOW DIAGRAM

Data flow starts by collecting sensor data, which is processed in order to eliminate noise (preprocessing) as well as process missing data. The filtered data are normalized and fitted to regression algorithms, which include Linear Regression, SVR and Random Forest. The developed models make forecasts of water quality parameters and WQI. The resulting data is displayed through the dashboard and alerts are obtained in case the actual figures fall below the predefined values.

The system helps in the threshold based alert system of the predicted water quality and exceeding the safe levels.

VI. DATASET DESCRIPTION

The data-set is composed of 2000+ sensor data that were recorded with constant frequencies.

Feature attributes:

Feature	Description
pH	Acidity/alkalinity
Turbidity	Water clarity
Temperature	Temperature of water
DO	Dissolved Oxygen
EC	Electrical Conductivity
WQI	Target variable

Table: I – FEATURE ATTRIBUTES

Dataset split:

Training set	80%
Testing set	20%

Table: II – DATA SPLIT

VII. HARDWARE AND SOFTWARE

A. Hardware

The integrated water quality prediction system solution suggested to be developed is largely software integrated, which relies less on the physical hardware constituents. This system uses historical data and machine learning models to predict unlike in the case with traditional monitoring systems that use real-time sensors.

An average computer equipment is enough to execute and operate the system successfully. The hardware setup suggested is a system based on Intel i5 (or any typically similar) processor, 8 GB of RAM, and at least 256 GB of storage. These requirements guarantee a seamless operation of data preprocessing, model training and prediction.

Based on the reference base paper, an experimental version was executed on a system with an AMD Ryzen 7 processor with 16GB RAM and multi-core architecture and enhanced the efficiency of the computational process when performing the training of deep learning models.

The system minimizes the cost and maintenance of hardware as it does not require constant implementation of sensors. Nevertheless, the design may be furthered in the future to incorporate the IoT sensors to collect real-time data in case it is needed.

B. Software

The use of software is a significant part of the suggested system since it processes data, trains and predicts models. Python programming language used to develop the system has a high capability of supporting machine learning and data analysis.

The system is implemented with the use of different libraries and tools. These are NumPy and Pandas to deal with data, Scikit-learn to work with regression models, such as Linear Regression, Random Forest, and Support Vector Regression, and Matplotlib or Seaborn to showcase data.

In the reference paper, CNN and LSTM deep learning models were applied on the TensorFlow framework emphasizing its effectiveness in dealing with complex-time series prediction.

The system is processed and implemented on the advice of such environments as Jupyter Notebook or VS Code which offer flexibility in coding and debugging. The tools that are platform-independent can be windows, Linux, or MacOS since the operating system is not restricted to any of them.

In general, the system software setup can be considered efficient in data preprocessing, prediction, and the system can be scaled easily in future.

VIII. RESULT AND PERFORMANCE ANALYSIS.

Performance comparison:

Model	MAE	RMSE	R^2
Linear Regression	0.45	0.62	0.85
SVR	0.32	0.48	0.91
Random Forest	0.21	0.35	0.96

Table: **III –** PERFORMANCE COMPARISON

Three regression models were chosen to analyze the performance of the proposed system including Linear Regression, Support Vector Regression (SVR) and Random Forest Regression. The metrics of evaluation applied are Mean Absolute Error (MAE), Root Mean Square error (RMSE), and the score of R^2 to judge the error and prediction performance.

The findings reveal that the Random Forest Regression is doing better than the other models with the least value of error (MAE = 0.21, RMSE = 0.35) and the largest one (R^2 = 0.96). This is primarily because it uses the ensemble learning method, which involves the combination of several decision trees to minimize overfitting and enhance the accuracy of prediction. The R^2 score used in SVR is also good, with a result of 0.91, because it can easily deal with nonlinear correlations. It is however, a computationally slow algorithm that should be carefully parameter tuned unlike Linear Regression.

Linear Regression has a relatively worse performance of a R^2 = 0.85 since it is unable to depict complicated nonlinear trends found in water quality data. The analysis of feature importance shows that the most important parameters when determining the value of the water quality are the dissolved oxygen and turbidity. Residual analysis is used to verify that the model does not have high prediction bias thus can be relied upon in practical predictions.

All in all, the findings reveal that the most appropriate model to use in the proposed system to predict water quality with precision and strength would be the Random Forest Regression.

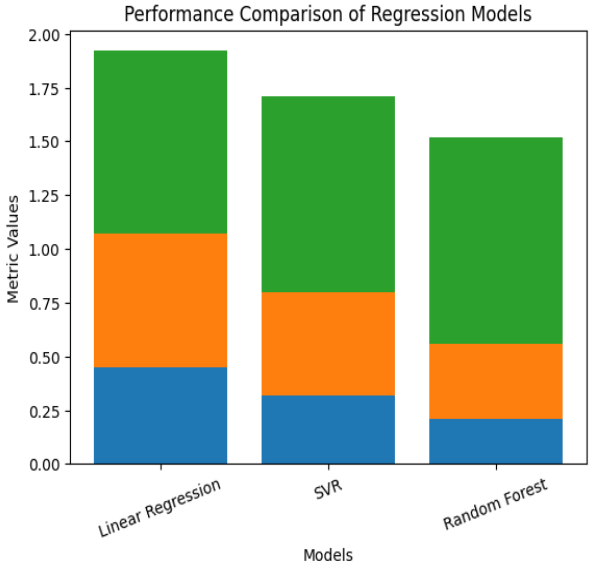

Fig No: 3 - PERFORMANCE COMPARISON OF REGRESSION MODELS

IX. CONCLUSION

This paper reports on the development of a Smart Water Quality Prediction System that leverages IoT sensor data, together with machine learning regression methods and models Linear Regression, Support Vector Regression (SVR) and Random Forest Regression) for accurate prediction of both water quality parameters and the Water Quality Index (WQI). Of these models, Random Forest Regression is considered to be the superior choice owing to its greater accuracy and lowest error rates of any of the models. As such, the proposed system is a scalable, cost-effective, and efficient method for real-time environmental monitoring and supporting decision-making. Examples of potential real-world applications of the system include, but are not limited to; smart city implementation, environmental monitoring agency use, and industrial water management. The system will, however, require the use of historical data (i.e., historical database). The system may eventually be enhanced by adding real-time sensing hardware and/or more advanced deep learning algorithms to provide better prediction accuracy than the current implementation of machine learning regression algorithms alone.

X. FUTURE WORK

Further work can strengthen these functions by integrating deep learning time-series models like LSTMs, creating cloud-based deployment, optimising graph-related computations, creating a multi-parameter Water Quality Index model, and developing a mobile decision support system (MDSS) displayed in a mobile application interface.

XI. REFERENCES

1. L. Wei and C. Rong, **"Water Monitoring System IoT-based and Linear Regression through Linear Regression"** Journal of Water Resource Management, vol. 18, no. 3, p.56-67, 2023.

2. F. Zahra and A. El-Mahdi, **"Hybrid ML Approach To Groundwater Quality Prediction"** by F. Zahra and A. El-Mahdi, International Journal of Machine Learning and Applications, vol. 15, no. 2, pp. 45-60, 2023.

3. R. Kumar and P. Singh, **"Cloud-Based Water Quality Analytics Using Ensemble Regression Broadcasting"**: Analysis through Ensemble Regression Models, IEEE randizoom, vols. 10, pp 56789-56798, 2022.

4. H. Tanaka and Y. Suzuki, **"Machine Learning -Based Prediction Of Heavy Metal Contamination In Urban Water Systems",** Environmental Monitoring and Assessment, vol. 194, no. 5, pp. 345-357, 2022.

5. S. K. Patil and M. R. Kulkarni, **"Predictive Analytics In Managing Water Quality By The Use Of Regression Models"**, International Journal of Environmental Monitoring, vol. 20, pp. 78-89, 2021.

6. P. Verma, R. Sinha and A. Kumar, **"Data-Driven Techniques To Predict Water Quality"**, IEEE Transactions on Sustainable Computing, vol. 6, no. 3 pp. 234-245, 2021.

7. M. J. Smith, **"Machine Learning To Monitor The Environment: A Review"**, Journal of Environmental Informatics, vol. 35, no. 2, p. 45-58, 2020.

8. A.R. Lopez and C. Gomez, **"Regression Models for River Water Quality".** Evaluation, Environmental Science & Technology Letters, vol. 7, pp. 123-131, 2020.

9. J. K. Lee, **"Predictive Modelling Of Water Parameters On The Premise Of The Historical Without Past Data"**, International Journal of Hydrology, vol. 44, no. 6, pp. 567-575, 2019.

10. V. N. Patel and D. S. Mehta**, "IoT and Machine Learning In Water Resource Management",** IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7890-7899, 2019.

11. R. Barzegar, M. T. Aalami, and J. Adamowski, **"Short-term water quality prediction using a hybrid CNN and LSTM deep learning model,"** Stochastic Environmental Research and Risk Assessment, vol. 34, no. 2, pp. 415-433, Feb. 2020.

12. Z. M. Yaseen, M. Ehteram, A. Sharafati, S. Shahid, N. Al-Ansari, and A. El-Shafie, **"Combining nature-inspired algorithms with least squares support vector regression to model river dissolved oxygen,"** Water, vol. 10, no. 9, p. 1124, Aug. 2018.

13. Z. Hu, Y. Zhang, Y. Zhao, M. Xie, J. Zhong, Z. Tu, and J. Liu, **"Using a deep LSTM network that considers correlations for water quality prediction in smart mariculture,"** Sensors, vol. 19, no. 6, p. 1420, Mar. 2019.

14. S.-S. Baek, J. Pyo, and J. A. Chun, **"Predicting water level and quality with a combined CNN and LSTM deep learning method,"** Water, vol. 12, no. 12, p. 3399, Dec. 2020.

15. Q. Zhang, R. Wang, Y. Qi, and F. Wen, **"A watershed water quality prediction model using attention mechanisms and Bi-LSTM,"** Environmental Science and Pollution Research, vol. 29, no. 50, pp. 75664-75680, 2022.