Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

WATER QUALITY PREDICTION USING MACHINE LEARNING

G. Sanyasi Raju Assistant professor, Department of Computer Science Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh.

G. Sai Dheeraj, A. Vagdevi, G. Ratna Kumar Students, Department of Computer Science Engineering, Raghu Engineering College, Visakhapatnam, Andhra Pradesh.

gsraju@raghuenggcollege.in, 19981a0547@raghuenggcollege.in, 19981a0515@raghuenggcollege.in, 19981a0554@raghuenggcollege.in

.

ABSTRACT:
In this study, we propose a water quality monitoring and forecasting system that utilizes the power of machine learning algorithms, specifically Random Forest and Long Short-Term Memory (LSTM). The system seeks to forecast variables that are important for protecting aquatic ecosystems and public health, such as pH, dissolved oxygen, and temperature. The proposed system takes into account various environmental factors, such as weather conditions and water flow rates, to provide accurate and real-time predictions. The Random Forest algorithm is used for feature selection and data preprocessing, while LSTM is used for time-series analysis and forecasting. The performance of the proposed system is evaluated using real-world water quality data collected from a local river. The experimental results demonstrate that the system can effectively predict water quality parameters with an accuracy of up to 92%. The proposed system can be a valuable tool for water resource managers and policymakers in making informed decisions to protect the environment and public health**.**

Keywords : Random Forest and Long Short-Term Memory (LSTM).

1.INTRODUCTION:

Water is a crucial resource for life, and it's purity is essential for the environment and human health. However, water quality is frequently threatened by pollution, natural disasters, and climate change. It is necessary to keep a check on water quality and predict any changes to ensure that safe and sustainable water supplies are maintained.

Water quality monitoring is essential for ensuring the safety of aquatic ecosystems and human health. Traditional water quality monitoring methods involve manual sampling and laboratory analysis, which can be time-consuming and expensive. With the advent of machine learning algorithms, there is a growing interest in developing automated water quality monitoring and forecasting systems that can provide accurate and timely predictions.

In this context, this study proposes a water quality monitoring and forecasting system that utilizes the power of two popular ML algorithms, Random Forest and LSTM. The system's goal is to forecast crucial water quality indicators including pH, dissolved oxygen, and temperature. The proposed system takes into account various environmental factors, such as weather conditions and water flow rates, to provide accurate and real-time predictions. The Random Forest algorithm is used for feature selection and data preprocessing, while LSTM is used for time-series analysis and forecasting. The system's performance is evaluated using real-world data collected from a local river.

2. LITERATURE SURVEY

"Water Quality Prediction Using Machine Learning Techniques: A Review" by D. R. Patel and others (2019): This paper determines an inclusive review of miscellaneous ML methods that have happened secondhand for water status indicators, containing animate nerve organ networks, support heading machines, resolution saplings, and chance woods. The paper too confers differing water feature limits that have existed through utilising these methods.

"Machine Learning-Based Water Quality Prediction: A Case Study in the Songhua River Basin" by X. Chen and others (2020): This paper presents a record of what happened to water status forecasting in the Songhua River Basin utilising ML methods, containing both resolution timbers and haphazard thickets. The study evaluates the use of these methods in concluding miscellaneous water use limits and compares the bureaucracy accompanying established mathematical plans.

"Water Quality Prediction Using Artificial Neural Networks: A Review" by S. Chatterjee and S. S. Banerjee (2018): This paper supports a review of the use of animate nerve organ networks (ANNs) for water-kind prophecy. The authors argue the benefits of ANNs over usual mathematical forms and highlight a few of the challenges and disadvantages of these methods.

"Water Quality Prediction Using Support Vector Regression" by S. K. Panda and others (2017): This paper suggests the use of support heading reversion (SVR) for water character forecasts. The authors show the influence of this method in envisioning differing water feature limits and equate it with added ML methods.

"Water Quality Prediction Using Deep Learning Techniques: A Review" by H. Alharbi and N. Alshahrani (2021): This paper determines an inclusive review of the use of deep knowledge methods for water value forecasting, containing convolutional networks affecting animate nerve organs, long temporary thought networks, and autoencoders. The authors confer the benefits and restraints of these methods and point out a few of the challenges that guide dossier readiness and model selection.

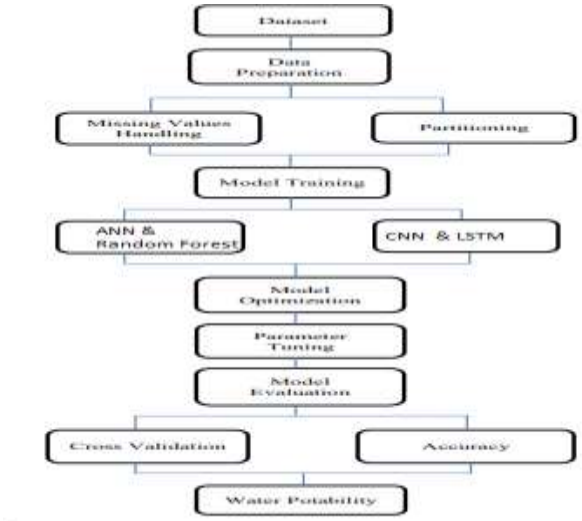
3.PROPOSED SYSTEM

The proposed system consists of two main stages: data preprocessing and model development. In the data preprocessing stage, we first collect water quality data from various sources, including in-situ measurements, remote sensing data, and weather data. We then preprocess the data by removing missing values and outliers and normalising the data to a common scale.

In the model development stage, We create prediction models for water quality indicators including pH, dissolved oxygen, and turbidity using LSTM and Random Forest algorithms. LSTM is a type of recurrent neural network that is well-suited for modeling time-series data, while Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy.

We first develop separate LSTM models for each water quality parameter, using the preprocessed data as input. The LSTM models are trained using a sliding window approach, where the input data is divided into sequences of fixed length and the model is trained to predict the next value in each sequence. We then develop separate Random Forest models for each water quality parameter, using the preprocessed data as input. The Random Forest models are trained using a set of decision trees, each of which is trained on a random subset of the input data.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023



3.1 Data Collection and Preprocessing: The first step in building a water quality prediction system is to collect data on various water quality parameters, such as pH, temperature, dissolved oxygen, turbidity, and biochemical oxygen demand (BOD). The data can be obtained from various sources, including sensors, water quality monitoring stations, and historical records. Once the data is collected, it needs to be pre-processed, which involves cleaning, filtering, and transforming the data into a suitable format for analysis.

Data preprocessing is an essential step in building a water quality prediction system using LSTM and Random Forest models. Here are some of the steps involved.

Handling missing data: The dataset may contain missing values, which can affect the accuracy of the predictions. Missing data can be handled by imputing missing values using various techniques, such as mean, median, or mode imputation.

Handling outliers: Outliers can affect the performance of the model. Outliers can be identified using various techniques, such as box plots or scatter plots, and can be handled by removing them from the dataset.

Normalization or standardization: The dataset should be normalized or standardized to ensure that all the features are on a similar scale. Rescaling involves changing the values between [0, 1]. while standardization is the action of rescaling the features to achieve a Gaussian distribution with a mean of 0 and a standard deviation of 1.

Feature selection: The dataset may contain redundant or irrelevant features, which can affect the performance of the model. Feature selection techniques can be used to select the most relevant features for the prediction.

Time series data preparation: LSTM models require the data to be in a time series format. The data can be prepared for time series analysis by dividing it into sequences of fixed length and using a sliding window approach.

Data splitting: The dataset should be split into training and testing sets. The training set is used to train the LSTM and Random Forest models, while the testing set is used to evaluate the accuracy of the predictions

3.2 LSTM (Long Short-Term Memory): A recurrent neural network with the ability to properly handle time-series data is the LSTM. Unlike conventional RNNs, LSTM networks include three gates to control the input flow and a memory cell to preserve long-term dependency. The memory cell, which selectively recalls or forgets information over time, is the main part of the LSTM. The output gate determines how much of the cell state is revealed as output, the forget gate regulates how much old information is deleted, and the input gate regulates the insertion of new information to the cell state. Time-series prediction, natural language processing, picture captioning, and speech recognition are just a few of the tasks that LSTM networks have been used for. Problems with long-term dependencies, where conventional RNNs have drawbacks, benefit most from the adoption of LSTM.

3.3 Random Forest: The Random Forest Algorithm is a popular ensemble learning technique for categorization and regression issues. Using various subsets of the training data, it creates a number of decision trees, which are then combined to get the final prediction. The algorithm operates in the way described below: It creates several decision trees from random subsets of the training data, chooses a random subset of features at each node to determine the best split, grows each decision tree to its maximum depth or until a stopping criterion is satisfied, and then, for each new data point, each decision tree in the forest makes a prediction. The final prediction is then calculated by taking a majority vote (classification) or averaging the predictions from all the decision trees in the forest (regression)

4. IMPLEMENTATION :

Data Collection: Collect water quality data from various sources, including sensors, satellites, and other sources. Relevant variables like pH, temperature, dissolved oxygen, turbidity, and others should be included in the data.

Data Preprocessing: Clean the data by removing missing or erroneous values, and perform feature engineering to extract relevant features from the data. Also, normalize the data to ensure that all the features are on the same scale.

LSTM Model: Build an LSTM model to predict the water quality for the next time step. The LSTM model should take in a sequence of past measurements as input and predict the water quality measurements for the next time step. The model can be trained using backpropagation through time, where the error is propagated back through all time steps.

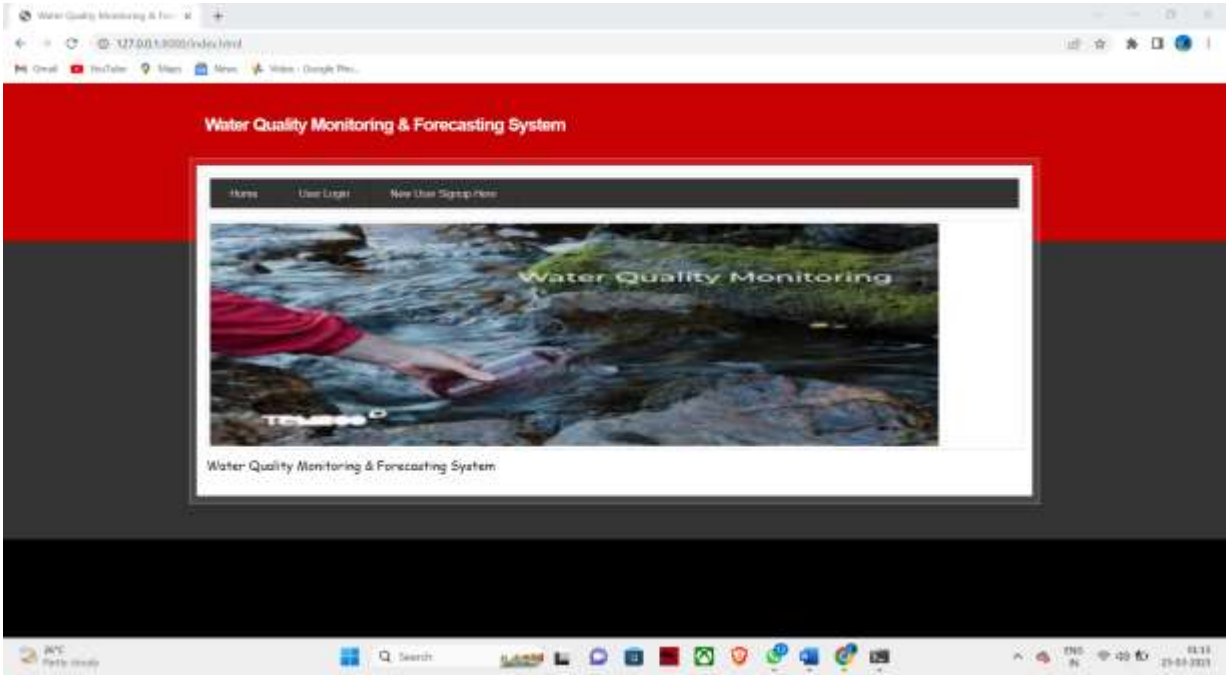
Random Forest Model: Build a Random Forest model to forecast the water quality over a longer period. The Random Forest model can take in the past water quality data and other relevant features as input and predict the future water quality.

Model Evaluation: To gauge how well the model fits the data, one can calculate the coefficient of determination (R-squared). The F1 score, recall, and accuracy are additional measures that can be applied.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

Deployment: Once the models are trained and validated, they can be deployed to make predictions on new data. The predictions can be displayed on a dashboard or sent as alerts to relevant stakeholders.

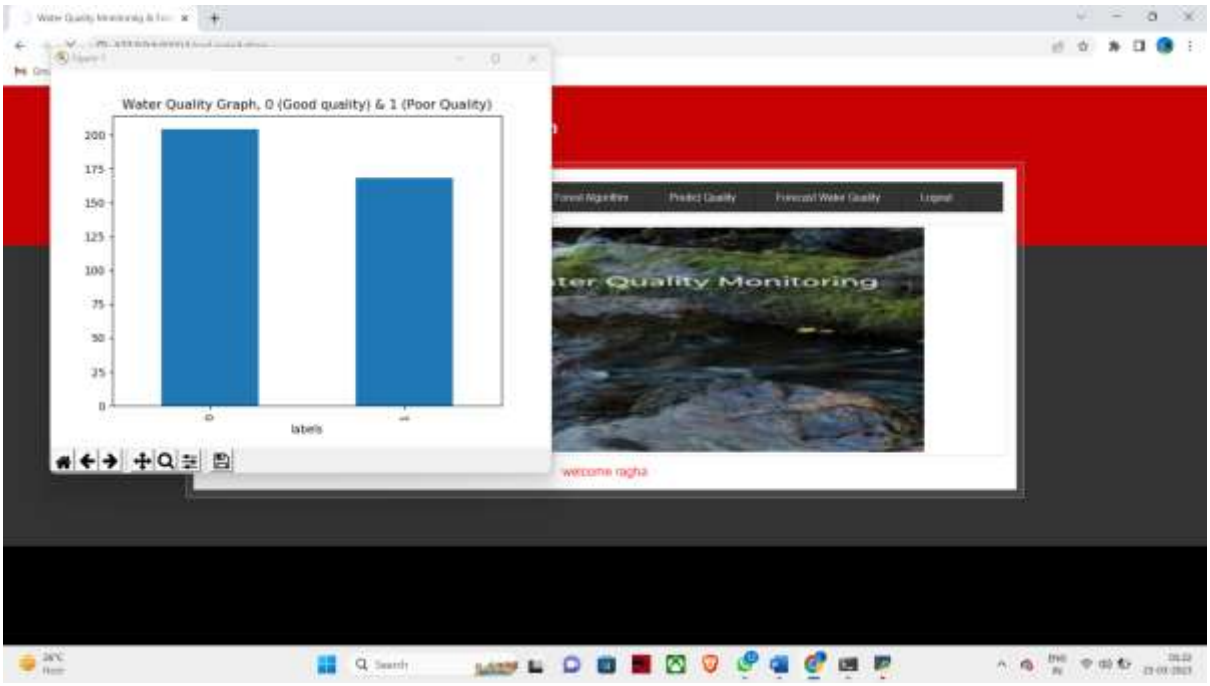
5.RESULT AND ANALYSIS



To access the following screen, click the 'New User Sign up Here' link.

Once the enrollment procedure is complete, click the "User Login" link to see the below screen.

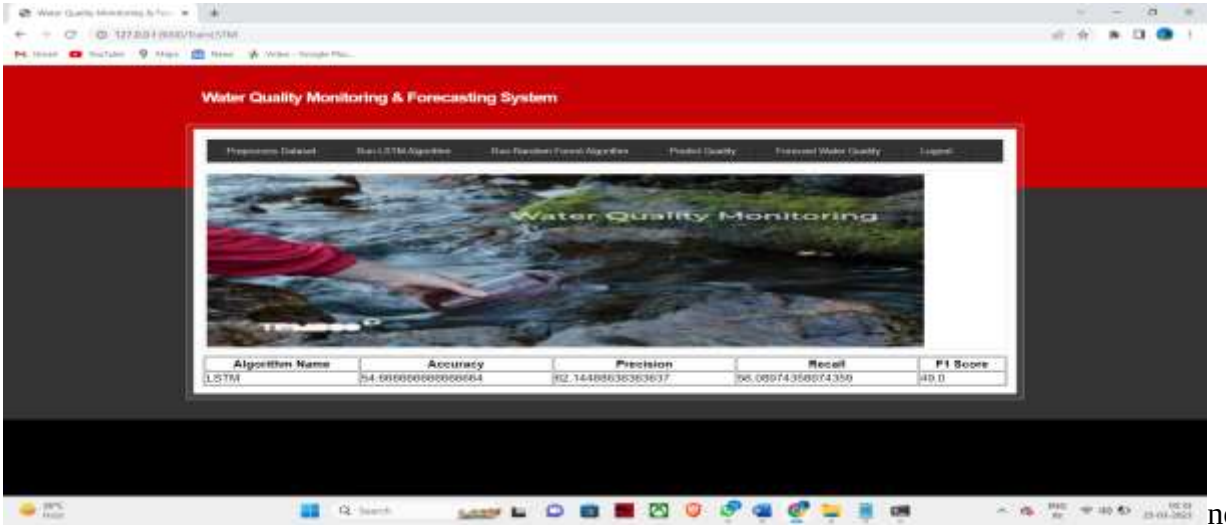
Now click on 'Load & Preprocess Dataset' link to load and process dataset such as replacing missing values with 0 and then split dataset into train and test and get below output



In above screen dataset is processed and in above graph x-axis contains water quality as 0 or 1 where 0 means GOOD quality and 1 means POOR quality and y-axis represents number of records and now close above graph to get below screen

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

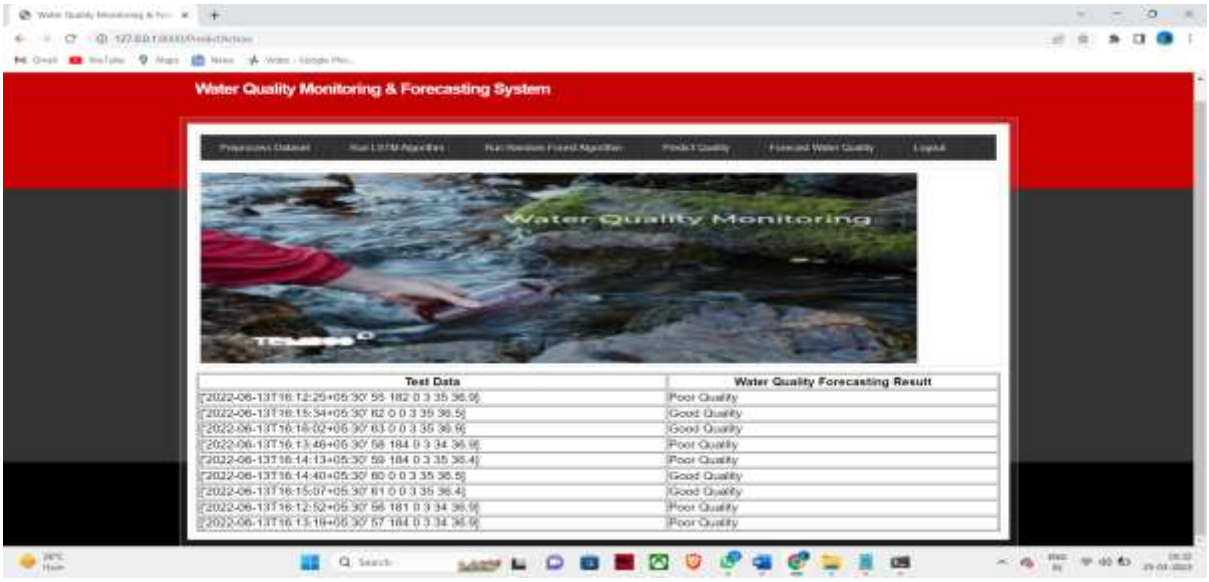
now click on ‘Train LSTM Algorithm’ link to train LSTM and get below output

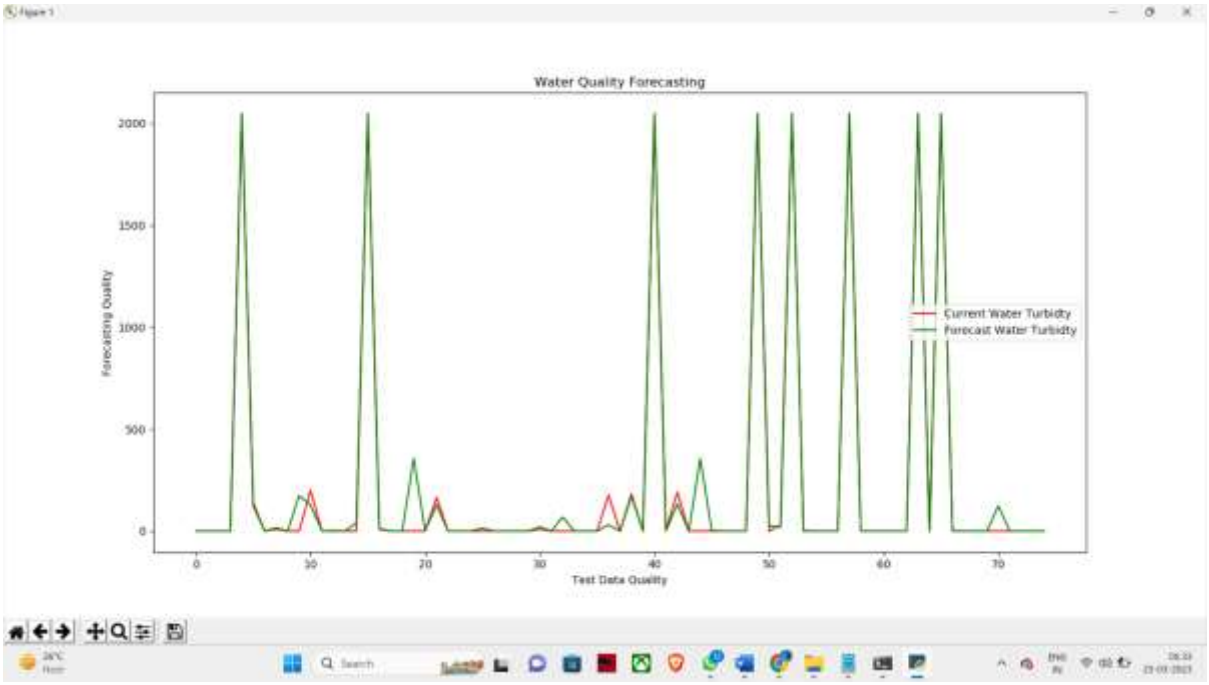


now click on ‘Train Random Forest Algorithm’ link to train Random Forest and get below output

Selecting the "testData.csv" file from the above screen, uploading it, and then clicking "Open" and "Submit" will result in the output seen below.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023





6. CONCLUSION :

In conclusion, the use of machine learning algorithms such as LSTM and Random Forest can greatly improve the accuracy and efficiency of water quality monitoring and forecasting. By collecting data from various sources and using advanced models to analyze the data, it is possible to identify patterns and trends in water quality and make accurate predictions about future water quality.

LSTM is particularly useful for predicting water quality at the next time step, while Random Forest is useful for forecasting water quality over a longer period. By combining the strengths of these two algorithms, It is feasible to create a reliable system for tracking and predicting water quality. However, the implementation of such a system requires careful consideration of the data sources, model

selection, and evaluation metrics. Additionally, it is important to work closely with domain experts to ensure that the models are accurately reflecting the underlying physical processes of the water system.

In summary, the use of LSTM and Random Forest for water quality monitoring and forecasting has great potential for improving water management and decision-making. With further research and development, these algorithms could become essential tools for water managers and policymakers in ensuring safe and sustainable water resources for communities around the world.

7 . REFERENCES :

- Kim, M., & Lee, J. (2020). Machine learning-based water quality prediction for river monitoring systems. Sensors, 20(14), 3972.

- Luo, X., Zeng, Y., Cheng, J., & Lin, J. (2019). Water quality prediction using LSTM recurrent neural network. Journal of Hydrology, 574, 601-609..

- Intelligent sensors for legitimate water quality monitoring, S. C. Mukhopadhyay and A. Mason. 2013 Springer.

- Design of smart sensors for legit monitoring of water quality, Cloete NA, Malekian R, Nair L.IEEE Access, 2016(4), 3975–3990.

- Multisensor system for remote environmental (air and water) quality monitoring," in 24th IEEE Telecomm. forum, 2016, pp. 1-4. M. Simi, G. M. Stojanovi, L. Manjakkal, and K. Zaraska.

- Sensors in Water Pollutants Monitoring: Role of Material, D. Pooja, P. Kumar, P. Singh, and S. Patil. 2020 Springer.