VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belgavi-590 018, Karnataka, India

A DISSERTATION REPORT

On

Predictive Modelling of Urban Water Quality

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By

Siddhi Narayan	**USN: 1MV20CS108**
Dharmitha V	**USN: 1MV20CS122**
Y Yasaswini	**USN: 1MV20CS128**

Carried out at
Department of CSE,
Sir M. Visvesvaraya Institute of Technology

Under the guidance of
Dr. Suma Swamy
Professor
Department of CSE
Sir MVIT

SIR M. VISVESVARAYA INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
BENGALURU-562157

2023-2024

SIR M. VISVESVARAYA INSTITUTE OF TECHNOLOGY
Bengaluru - 562157
Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project work entitled **"**Predictive Modelling of Urban Water Quality**"** is a bonafide work carried out by **Siddhi Narayan (1MV20CS108), Dharmitha V (1MV20CS122) and Y Yasaswini (1MV20CS128)** in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi** during the year **2023-2024.** It is certified that all corrections/suggestions indicated for internal assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements with respect to project work prescribed for the Bachelor of Engineering degree.

.....	.....	.....
Signature of Guide	Signature of HOD	Signature of Principal
Dr. Suma Swamy	**Dr. Anitha.T.N**	**Prof. Rakesh S G**
Professor	Professor & HOD,	Principal SMVIT,
Dept. of CSE, SMVIT	Dept. of CSE, SMVIT	Bengaluru

External Examiners:

Name of the Examiners	**Signature with Date**
1.	
2.	

DECLARATION

We **Siddhi Narayan, Dharmitha V and Y Yasaswini** student of VIII semester B.E in Computer Science and Engineering at Sir M. Visvesvaraya Institute of Technology, Bengaluru, hereby declare that this dissertation work entitled **"Predictive Modelling of Urban Water Quality"** has been carried out at Department of CSE, Sir M. Visvesvaraya Institute of Technology under the guidance of guide **Dr. Suma Swamy, Professor, Department of CSE, Sir M. Visvesvaraya Institute of Technology, Bengaluru** and submitted in the partial fulfilment for the award of degree Bachelor of Engineering in Computer Science and Engineering of Visvesvaraya Technological University, Belagavi during the academic year 2023-2024. We further declare that the report had not been submitted to any other university for the award of any other degree.

Place: Bengaluru
Date: 25-05-2024

SIDDHI NARAYAN
USN: 1MV20CS108

DHARMITHA V
USN: 1MV20CS122

Y YASASWINI
USN: 1MV20CS128

ABSTRACT

Urban water bodies face ongoing threats from pollution, necessitating advanced monitoring and prediction strategies to ensure timely intervention and resource preservation. This study explores the application of machine learning techniques for predictive modelling of water quality parameters in urban environments. Leveraging sensor data, satellite imagery, and historical records, the proposed models aim to forecast crucial indicators such as pH levels, dissolved oxygen, turbidity, and pollutant concentrations. The research emphasizes the development of accurate and robust machine learning algorithms capable of real-time monitoring, enabling early detection of potential contamination events.

By integrating predictive analytics into water quality management systems, this approach facilitates proactive decision-making, contributing to the sustainable preservation of urban water resources. The findings hold significant implications for environmental monitoring, policy formulation, and the creation of smart city infrastructures aimed at safeguarding water quality in the face of growing urbanization and environmental challenges.

This research delves into the realm of water quality prediction in urban settings by harnessing the power of machine learning (ML). Integrating data from diverse sources, including sensors and satellite imagery, the study focuses on creating predictive models for key water quality parameters such as pH, dissolved oxygen, turbidity, and pollutant concentrations. By employing advanced ML algorithms, the objective is to enable real-time monitoring and early identification of potential pollution events, thereby empowering decision-makers with actionable insights for effective water quality management. This innovative approach not only enhances environmental monitoring practices but also contributes to the development of smart city infrastructures capable of ensuring the sustained health of urban water bodies amidst the challenges of rapid urbanization and environmental changes.

ACKNOWLEDGEMENT

It gives us immense pleasure to express our sincere gratitude to the management of **Sir M. Visvesvaraya Institute of Technology,** Bengaluru for providing the opportunity and the resources to accomplish our project work in their premises.

On the path of learning, the presence of an experienced guide is indispensable and we would like to thank our guide **Dr. Suma Swamy,** Professor, Dept. of CSE, for her invaluable help and guidance.

Heartfelt and sincere thanks to **Dr. T. N. Anitha,** HOD, Dept. of CSE, for his suggestions, constant support and encouragement.

We would also like to convey our regards to **Prof. Rakesh S G,** Principal, Sir. MVIT for providing us with the infrastructure and facilities needed to develop our project.

We would also like to thank the staff of Department of Computer Science and Engineering and lab-in-charges for their co-operation and suggestions. Finally, we would like to thank all our friends for their help and suggestions without which completing this project would not have been possible.

Siddhi Narayan (1MV20CS108)
Dharmitha V (1MV20CS122)
Y Yasaswini (1MV20CS128)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER - 1
INTRODUCTION

1.1 Overview

Water quality is a critical aspect of environmental health and human well-being, as it directly impacts ecosystems, agriculture, and human consumption. With the increasing pressures on water resources due to population growth, industrialization, and climate change, ensuring water quality has become a paramount concern. Traditional methods of water quality monitoring involve manual sampling and laboratory analysis, which are often time-consuming and costly.

In recent years, there has been a growing interest in leveraging machine learning (ML) techniques to predict and monitor water quality more efficiently. ML algorithms can analyze large datasets containing diverse water quality parameters, environmental variables, and historical records to make accurate predictions about water quality conditions. This approach offers the advantage of real-time or near-real-time monitoring, allowing for timely interventions and proactive management of water resources.

The utilization of machine learning (ML) for water quality prediction offers a plethora of benefits that address the challenges inherent in traditional monitoring methods. Firstly, ML algorithms enable the early detection of contamination events, providing a crucial advantage in swiftly implementing preventive measures to safeguard public health and environmental integrity. Moreover, the predictive modeling capabilities of ML empower stakeholders to anticipate changes in water quality by analyzing historical data alongside relevant environmental variables. This not only aids in proactive management but also facilitates optimized resource allocation, ensuring that interventions are strategically deployed where they are most needed.

1.2 Problem Statement

Develop a system to predict water quality, addressing challenges in real-time data collection, preprocessing, visualization, and classification based on environmental factors like pH, dissolved oxygen (DO), Biochemical Oxygen Demand (BOD), Total Suspend Solids (TSS) and Nitrate- Nitrogen (NO3-N) etc. The primary focus aims to provide accurate assessments for effective water resource management and environmental monitoring with the objective of enhancing accuracy and efficiency in water quality assessment. Through the utilization of ML techniques, the Outcome of the model classifies the input data set in to potable or not-potable class.

1.3 Significance and Relevance of Work

- Early Detection of Contamination: ML models can analyze historical water quality data to detect patterns and anomalies. Early detection of changes in water quality can help identify potential contamination issues before they become severe, allowing for timely intervention and mitigation.
- Public Health Protection: Predictive models can help safeguard public health by providing warnings about potential waterborne diseases or contaminants. Timely alerts enable authorities to take preventive measures, such as issuing advisories or implementing water treatment processes.
- Resource Optimization: ML algorithms can optimize the allocation of resources for water treatment and distribution. By predicting water quality, utilities can adjust treatment processes more efficiently, reducing costs and ensuring that resources are used effectively.
- Environmental Monitoring: Water quality prediction contributes to environmental monitoring by assessing the impact of human activities on aquatic ecosystems. Identifying changes in water quality helps in understanding environmental trends, supporting conservation efforts, and preventing damage to ecosystems.
- Environmental Sustainability: With increasing concerns about environmental sustainability, predicting water quality becomes crucial for monitoring and mitigating the impact of human activities on aquatic ecosystems. This project contributes to sustainable water management practices.
- Climate Change Adaptation: As climate change continues to affect weather patterns and water sources, predicting water quality becomes essential for adapting to these changes. ML models can help in understanding and responding to the dynamic nature of water quality in the face of climate variability.
- Infrastructure Planning and Optimization: Water quality prediction supports the planning and optimization of water treatment and distribution infrastructure. This is particularly relevant as urbanization and population growth place increased pressure on water resources and the associated infrastructure.

1.4 Objectives

- **Design a Comprehensive ML Model:** Develop a machine learning model that integrates multiple algorithms to predict water quality accurately.

- **Implement a Classification System:** Input dataset will be classified to a particular class based on their characteristics and also different environmental factors and attributes.

- **Implement Data Visualization Modules:** Develop visualization modules to present water quality data in an insightful and comprehensible manner, aiding in better understanding and decision-making.

1.5 Methodology

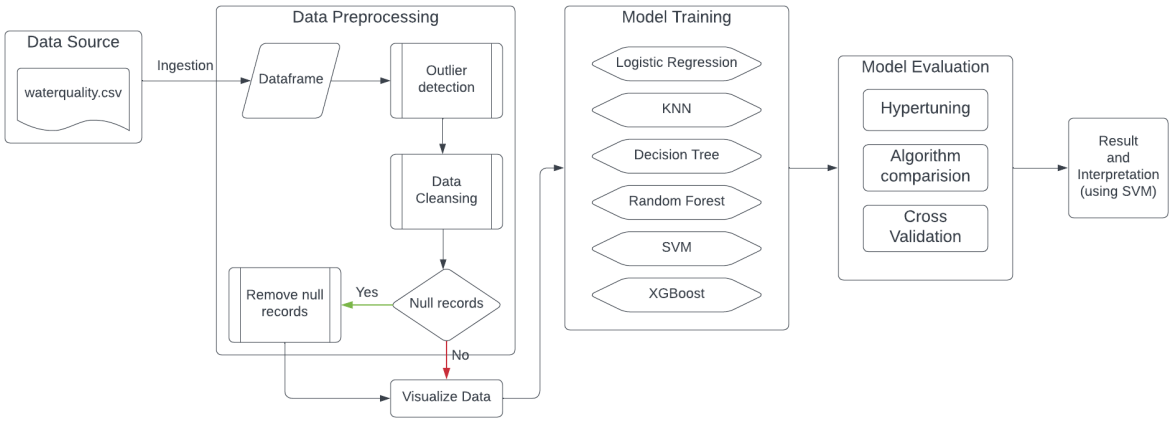


Fig 1.1 Methodology

1. **Data Preprocessing:** It is a crucial step in preparing raw data for analysis and modeling. It involves creating a data frame from a CSV file and performing data cleansing, which includes handling outliers and null records.
   - Outlier Detection: The data is scanned for outliers, which are data points that fall outside the expected range.
   - Data Cleansing: This step involves removing or correcting any errors in the data, such as null values. Here, the process checks for null records and removes them if there are any.
2. **Data Visualization:** The distributions of each dimension in the dataset were visualized using histograms with KDE plots, while outliers were identified through box plots. Additionally, a heat map was employed to illustrate the correlations among all dimensions, providing a comprehensive overview of the data's relationships and anomalies.
3. **Model Training:** Various classification models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost and K-Nearest Neighbors, were trained on the water quality dataset. Their accuracies were evaluated, and confusion

matrices were used to summarize the performance of each model, providing insights into their predictive accuracy and classification performance.

4. **Model evaluation:** It is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.
    - Hyper tuning: Training your model sequentially with different sets of hyperparameters.
    - Cross-validation: Evaluating models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.
5. **Result:** Post prediction analysis is discussed to finalize our results and review any recommendations with our final conclusions with intent to deploy a robust water quality classifier.

1.6 Organization of the Report

The report begins with an executive abstract, providing a brief overview of the project objectives, methodology, key findings, and conclusions.

Following the abstract, the report includes an introduction section that establishes the background and significance of water quality prediction and outlines the project objectives.

A comprehensive literature review is presented, summarizing existing research and methodologies related to water quality prediction and similar projects.

The methodology section details the data collection process, including sources, variables, and preprocessing steps. It also explains the feature selection techniques and machine learning models employed.

Results are presented, showcasing the performance metrics of the developed models, supported by visualizations such as graphs or tables to enhance understanding. A thorough discussion section analyzes and interprets the results in relation to existing literature, discusses strengths and limitations of the predictor, and suggests avenues for future improvement.

The report concludes by summarizing the key findings, achievements, and implications of the water quality prediction project, reiterating the project objectives and assessing the overall success and impact.

Finally, a list of references is provided, citing all sources used in the report, and appendices is included for supplementary information such as code snippets, sample data, or technical details.

CHAPTER – 2

LITERATURE SURVEY

2.1 Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction [IEEE, Volume:9] [26 July 2021]

Authors: Ali Omran Al-Sulttani, Mustafa Al-Mukhtar, Ali B. Roomi, Aitazaz Ahsan Farooque, Khaled Mohamed Khedher, Zaher Mundher Yaseen

This study developed five ensemble machine learning models—Quantile Regression Forest (QRF), Random Forest (RF), Radial Support Vector Machine (SVM), Stochastic Gradient Boosting (GBM), and Gradient Boosting Machines (GBM_H2O)—to predict monthly Biochemical Oxygen Demand (BOD) values in the Euphrates River, Iraq, utilizing ten years of monthly average data for various water parameters. The research compared the standalone models with integrated models that employed Genetic Algorithm (GA) and Principal Components Analysis (PCA) for feature extraction, assessing their effectiveness in predicting BOD values.

Disadvantages

- Complexity and Computational Cost
- Risk of Overfitting
- Data Quality and Availability
- Interpretability

2.2 Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality [IEEE Volume 11] [14 September 2023]

Authors: Mushtaque Ahmed Rahu, Abdul Fattah Chandio, Khursheed Aurangzeb, Sarang Karim, Musaed Alhussein, Muhammad Shahid Anwar

The paper presents an integrated IoT and machine learning framework for water quality analysis using sensors in the Rohri Canal, SBA, Pakistan. Machine learning models, including Support Vector Machine (SVM), XGBoost (eXtreme Gradient Boosting), Decision Trees, and Random Forest, were employed. Multilayer Perceptron (MLP) was used for regression to predict the Water Quality Index (WQI), while Random Forest was used for classification to predict Water Quality Class (WQC). The results indicate that MLP outperforms in regression tasks, and Random Forest excels in classification tasks. The study

also notes that the models perform better with smaller datasets, showing improved regression accuracy and superior classification metrics compared to existing studies. However, the study is limited by its use of a two-year dataset, restricting long-term trend analysis, and the exclusion of climate change variables, limiting comprehensive analysis. Future research should incorporate broader datasets, additional assessment metrics, and advanced models, and clarify the methodology's applicability to climate change analysis.

Disadvantages

- Limited long-term trend analysis
- Exclusion of climate change variables
- Need for broader datasets
- Requirement for additional assessment metrics

2.3 Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach [IEEE Volume 10] [10 November 2022]

Authors: Bilal Aslam, Ahsen Maqsoom, Ali Hassan Cheema, Fahim Ullah, Abdullah Alharbi, Muhammad Imran

The study aimed to improve Water Quality Index (WQI) predictions, which are typically time-consuming and error-prone, by utilizing machine learning. It employed four standalone algorithms—random trees (RT), random forest (RF), M5P, and reduced error pruning tree (REPT)—and 12 hybrid data-mining algorithms on well water samples from North Pakistan. The hybrid algorithms, which combined standalone, bagging, cross-validation, and randomizable filtered classification methods, showed superior performance compared to standalone models. The study suggests future research could examine algorithm performance over extended periods, incorporate key parameters like Chemical Oxygen Demand (COD) and Biochemical Oxygen Demand (BOD) over multiple years, and use deep learning algorithms such as convolutional neural networks for more comprehensive results. Additionally, applying tests like Principal Component Analysis (PCA) and investigating further water quality variables beyond correlation tests would enhance the study's robustness.

Disadvantages

- Limited to short-term data
- Exclusion of crucial parameters like COD and BOD
- Potential benefits of deep learning not explored
- Further validation tests like PCA not conducted

2.4 Water quality prediction using machine learning models based on grid search method [Multimedia tools and Applications Journal] [29 September 2023]

Authors: Mahmoud Y. Shams, Ahmed M. Elshewey, El-Sayed M. El-kenawy, Abdelhameed Ibrahim, Fatma M. Talaat Zahraa Tarek

The study focuses on enhancing water quality prediction using machine learning models, including Random Forest (RF), Extreme Gradient Boosting (Xgboost), Gradient Boosting (GB), and Adaptive Boosting (AdaBoost). It optimizes parameters for classification and regression models, including KNN, DT, SVR, and MLP, using grid search. The dataset comprises 7 features and 1991 instances, with preprocessing steps involving mean imputation and data normalization. For classification, the GB model achieves a good accuracy in predicting Water Quality Classification (WQC), while the MLP regressor excels in regression with a good R^2 value for predicting the Water Quality Index (WQI). However, the study has limitations, including potential lack of generalizability due to the specific dataset and geographic focus. Additionally, while grid search is used for parameter tuning, the models' sensitivity to these parameters may not be thoroughly explored. The chosen assessment metrics may not fully cover all performance aspects, and the study lacks an uncertainty analysis.

Disadvantages

- Limited generalizability
- Insufficient exploration of parameter sensitivity
- Incomplete performance assessment metrics
- Lack of uncertainty analysis

2.5 Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm [IEEE Volume 11] 2022

Authors: S. R. Sannasi Chakravarth, N. Bharanidharan, Vinoth Kumar Venkatesan, Mohamed Abbas, Harikumar Rajaguru

Water is essential for human survival and development, making water quality prediction crucial. The study introduces the Adaptive Crow Search Optimized SoftMax-Extreme Learning Machine (AdCSO-sELM) to enhance ELM performance by adaptively adjusting the flight length with iterations. The novelty lies in dynamically adapting the CSOA parameters, resulting in improved ELM performance. The AdCSO-sELM achieves a better accuracy in classifying water potability using the Kaggle dataset. However, limitations of the AdCSO-

sELM model include potential data dependency issues, reliance on specific datasets like Kaggle, and limited generalizability to diverse water quality contexts. The dynamic parameter adaptation increases complexity and computational overhead, which may affect real-time applications. Moreover, the model's performance can vary under different environmental conditions and requires extensive parameter tuning for optimal results.

Disadvantages

- Potential data dependency issues
- Reliance on specific datasets (e.g., Kaggle)
- Limited generalizability to diverse contexts
- Increased complexity and computational overhead
- Variable performance under different conditions
- Extensive parameter tuning required

2.6 Dual Kidney-Inspired Algorithm for Water Quality Prediction and Cancer Detection [IEEE Volume 8] 2020

Authors: Salwani Abdullah, Najmeh Sadat Jaddi

The Kidney-inspired Algorithm (KA) is a metaheuristic search algorithm modeled after the physiological processes of human kidneys. It mimics how the second kidney filters solutes if the first kidney fails, and dialysis as a backup if both fail, with Glomerular Filtration Rate (GFR) used to measure kidney function. The Dual-Kidney-inspired Algorithm (Dual-KA) extends this concept and has been applied to water quality prediction and cancer detection, demonstrating high effectiveness in these areas. However, the Dual-KA faces limitations such as direct biological mapping complexity, specificity to certain problems, challenges in interpretability, and scalability concerns.

Disadvantages

- Direct biological mapping complexity
- Specificity to certain problems
- Interpretability challenges
- Scalability concerns

2.7 A Complete Proposed Framework for Coastal Water Quality Monitoring System with Algae Predictive Model [IEEE Volume 9] 2021

Authors: Nur Aqilah Paskhal Rostam, Nurul Hashimah Ahamed Hassain Malim, Rosni Abdullah, Abdul Latif Ahmad

The utilization of Long Short-Term Memory (LSTM) networks provides an effective solution for predicting Harmful Algal Bloom (HAB) growth, crucial for managing water quality. Previous research has often focused separately on either the prediction aspect or the implementation of water monitoring systems that integrate sensor technology via the Internet of Things (IoT). In this context, LSTM has proven particularly adept, outperforming other basic machine learning methods in accurately predicting algal growth by tracking chlorophyll-a (Chl-a) levels, a key indicator of algal presence in coastal areas. This capability allows for the creation of a comprehensive, end-to-end framework for HAB prediction that incorporates both IoT-enabled data collection and sophisticated predictive modeling.

Disadvantages

- Algorithm adaptability challenges
- Sensor integration difficulties
- Data constraints and quality issues
- Generalization issues across different regions
- Limited interpretability of deep learning models

2.8 Machine learning algorithms for efficient water quality prediction [Modelling Earth Systems and Environments algorithm] 2021

Authors: Mourade Azrour, Jamal Mabrouki, Ghizlane Fattah

The Multiple Regression Algorithm leverages four specific water parameters—temperature, pH, turbidity, and coliforms—to predict the water quality index, demonstrating its effectiveness and importance in assessing water quality. While the algorithm simplifies the evaluation process by focusing on a limited set of parameters, this simplicity also enables more straightforward interpretations and applications in scenarios where these parameters are known to be critical indicators. However, the reliance on only a handful of factors might not capture the full complexity and variability of water quality dynamics, particularly in diverse environmental settings.

Disadvantages

- Oversimplification with limited parameters
- Variability in data reliability
- Challenges in interpretability with complex models
- Inability to capture dynamic environmental changes

2.9 Towards design of IoT and Machine Learning Embedded Frameworks for analysis and prediction of water quality [IEEE Volume 11] 2023

Authors: Mushtaque Ahmed Rahu, Abdul Fattah Chandio, Khursheed Aurangzeb, SarangKari, Musaed Alhussein

In this study, regression models including LSTM (Long Short-Term Memory), SVR (Support Vector Regression), MLP (Multilayer Perceptron), and NARNet (Nonlinear Autoregressive Network) are used to predict the Water Quality Index (WQI), while classification models like SVM (Support Vector Machine), XGBoost (eXtreme Gradient Boosting), Decision Trees, and Random Forest are used to predict Water Quality Classification (WQC). The ensemble approach, particularly in regression, shows potential in predicting water quality over time series data. However, this methodology comes with inherent limitations. The models' dependency on specific regressors might not adapt well to different datasets. Parameter settings heavily influence the performance, which could lead to varied outcomes if not optimally configured. Additionally, the generalizability of these models to different datasets, especially beyond the specific timeframe and geographic area of the study, remains questionable, indicating a need for further validation to ensure reliable application in broader contexts.

Disadvantages

- Limited adaptability to different datasets
- Heavy dependency on specific parameter settings
- Uncertain generalizability beyond the study's specific conditions
- Need for extensive validation for broader application

2.10 Dimensionality Reduction for Water Quality Prediction from a Data Mining Perspective [2020 Springer Nature Singapore Pte Ltd]

Authors: J. Alamelu Mangai, Bharat B. Gulyani

This study introduces a data-driven model utilizing SVM, MLP, linear regression (LR), and instance-based learner (IBK) to predict the Biochemical Oxygen Demand (BOD), a critical indicator of water quality. BOD measures the amount of dissolved oxygen required by aerobic organisms to break down organic material in water. Traditional methods for measuring BOD are often slow and costly. The proposed model addresses these issues by employing dimensionality reduction techniques like PCA and CFS to simplify the high-dimensional data, thereby enhancing the prediction process. However, the effectiveness of these techniques can vary based on the dataset's characteristics, potentially affecting the model's accuracy.

Disadvantages

- Potential suboptimality of PCA and CFS across various datasets
- Varying effectiveness depending on input feature characteristics
- Lack of detail on machine learning challenges like overfitting or outlier sensitivity
- Limited ability of evaluation metrics to gauge generalizability and robustness

Paper	**Author**	**Technique**	**Description**	**Limitations**
Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction [**IEEE, Volume:9**] [26 July 2021]	• Ali Omran Al-Sulttani • Mustafa Al-Mukhtar • Ali B. Roomi • Aitazaz Ahsan Farooque • Khaled Mohamed Khedher • Zaher Mundher Yaseen	Quantile regression forest (QRF), Random Forest (RF), radial support vector machine (SVM), Stochastic Gradient Boosting (GBM) and Gradient Boosting Machines (GBM_H2O)	This study developed five ensemble machine learning models (Quantile Regression Forest, Random Forest, Radial Support Vector Machine, Stochastic Gradient Boosting, and Gradient Boosting Machines) to predict monthly Biochemical Oxygen Demand (BOD) values in the Euphrates River, Iraq.	• Complexity and Computational Cost • Risk of Overfitting • Data Quality and Availability • Interpretability
Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality **[IEEE Volume 11**] [14 September 2023]	• Mushtaque Ahmed Rahu • Abdul Fattah Chandio • Khursheed Aurangzeb • Sarang Karim • Musaed Alhussein • Muhammad Shahid Anwar	SVM (Support Vector Machine), XGBoost (eXtreme Gradient Boosting), Decision Trees, and Random Forest	The paper proposes an integrated IoT and machine learning framework for water quality analysis, using sensors in Rohri Canal, SBA, Pakistan. Machine learning models, including MLP for regression and Random Forest for classification, predict Water Quality Index (WQI) and Class (WQC).	• Limited long-term trend analysis • Exclusion of climate change variables • Need for broader datasets • Requirement for additional assessment metrics

Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach **[IEEE Volume 10]** [10 November 2022]	• Bilal Aslam • Ahsen Maqsoom • Ali Hassan Cheema • Fahim Ullah • Abdullah Alharbi • Muhammad Imran	Random trees (RT), random forest (RF), M5P, and reduced error pruning tree (REPT)	This study aimed to enhance Water Quality Index (WQI) predictions, known for their time-consuming and error-prone nature, using machine learning. Four standalone algorithms (random trees, random forest, M5P, and reduced error pruning tree) and 12 hybrid data-mining algorithms were employed on well water samples from North Pakistan. Hybrid algorithms, combining standalone, bagging, cross-validation, and randomizable filtered classification methods, outperformed standalone ones.	Future work could explore algorithm performance over extended periods and include important parameters like COD and BOD over multiple years. Deep learning algorithms, particularly convolutional neural networks, could be employed for comprehensive results and comparison. Additionally, conducting tests like PCA and exploring further water quality variables beyond correlation tests would enhance the study's robustness.
Water quality prediction using machine learning models based on grid search method [**Multimedia tools and Applications Journal**] [29 September 2023]	• Mahmoud Y. Shams • Ahmed M. Elshewey • El-Sayed M. El-kenawy • Abdelhameed Ibrahim • Fatma M. Talaat Zahraa Tarek	Random forest (RF) model, Extreme Gradient Boosting (Xgboost) model, Gradient Boosting (GB) model, and Adaptive Boosting (AdaBoost)	This study focuses on improving water quality prediction using machine learning. It optimizes parameters for classification and regression models (RF, Xgboost, GB, AdaBoost, KNN, DT, SVR, MLP) through grid search. The dataset includes 7 features and 1991 instances. Preprocessing involves mean imputation and data normalization	The paper's limitations include a potential lack of generalizability due to its specific dataset and geographic focus. While grid search is used for parameter tuning, the sensitivity of models to these parameters may not be fully explored. The chosen assessment metrics for classification and regression models may not cover all aspects of performance, and the study lacks uncertainty analysis.

Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm **[IEEE Volume 11**] 2022	• S. R. Sannasi Chakravarth • N. Bharanidharan • Vinoth Kumar Venkatesan • Mohamed Abbas • Harikumar Raj aguru	Adaptive Crow Search Optimized SoftMax-Extreme Learning Machine (AdCSO-sELM)	Water is a predominant source in the survival and development of all human lives. On top of all, predicting water quality is a significant one since water is essential in regulating our human body. In recent days, the advent of machine learning techniques has been supporting a lot in water quality prediction. Accordingly, Adaptive Crow Search Optimized SoftMax-Extreme Learning Machine (AdCSO-sELM) is proposed to improve the ELM performance by making the flight length adaptively with respect to the iterations.	Limitations of the AdCSO-sELM model for water quality prediction include potential data dependency issues, reliance on specific datasets like Kaggle, and limited generalizability to diverse water quality contexts. The novelty in dynamic parameter adaptation might increase complexity and computational overhead, impacting real-time applications. Additionally, the model's performance might vary in different environmental conditions and require extensive parameter tuning for optimal results.
Dual Kidney-Inspired Algorithm for Water Quality Prediction and Cancer Detection **[IEEE Volume 8**] 2020	• Salwani Abdullah • Najmeh Sadat Jaddi	Dual-population Kidney-inspired Algorithm (Dual-KA)	The kidney-inspired algorithm (KA) was presented in a recent research paper as a metaheuristic search algorithm. The KA imitates the physiological process of the kidneys in the human body. The second kidney in the human body filters all the solutes if the other kidney fails. If the second kidney also gets damaged, dialysis can be performed as a treatment. The failure of a kidney is proved by the Glomerular Filtration Rate (GFR) calculation normal.	The Dual-Kidney-inspired Algorithm (Dual-KA) simplifies complex kidney processes for optimization but faces limitations in its direct biological mapping, specificity to certain problems, interpretability challenges, and scalability concerns.

A Complete Proposed Framework for Coastal Water Quality Monitoring System With Algae Predictive Model **[IEEE Volume 9]** 2021	• Nur Aqilah Paskhal Rostam • Nurul Hashimah Ahamed Hassain Malim • Rosni Abdullah; Abdul Latif Ahmad	Long Short-term Memory (LSTM)	An end-to-end process to achieve a complete framework methodology for Harmful Algal Bloom (HAB) growth prediction is crucial for water management. Previous works have separately focused on the prediction part or the implementation of the water monitoring system that involves the integration of sensors through the Internet of Things (IoT).	Creating an end-to-end Harmful Algal Bloom (HAB) prediction framework integrating IoT sensors and predictive modeling faces limitations in algorithm adaptability, sensor integration challenges, data constraints, generalization issues, and interpretability of deep learning methods.
Machine learning algorithms for efficient water quality prediction [Modelling Earth Systems and Environments algorithm] 2021	• Mourade Azrour, • Jamal Mabrouki, • Ghizlane Fattah	Multiple Regression Algorithm	Water is an essential resource for human existence. In fact, more than 60% of the human body is made up of water. Our bodies consume water in every cell, in the different organisms and in the tissues. Hence, water allows stabilization of the body temperature and guarantees the normal functioning of the other bodily activities. The method we propose is based on four water parameters: temperature, pH, turbidity and coliforms. The use of the multiple regression algorithms has proven to be important and effective in predicting the water quality index.	The approach's limitations include oversimplification with only four parameters, potential variability in data reliability, interpretability challenges and high computational demand with artificial neural networks, and the model's inability to capture dynamic changes in water quality influenced by external factors.

Towards design of IoT and Machine Learning Embedded Frameworks for analysis and prediction of water quality **[IEEE Volume 11**] 2023	• Mushtaque Ahmed Rahu • Abdul Fattah Chandio • Khursheed Aurangzeb • SarangKari • Musaed Alhussein	Support Vector Machine (SVM), XGBoost(eXtreme Gradient Boosting), Decision Tree and Random Forest	Regression models: LSTM (Long Short-Term Memory), SVR (Support Vector Regression), MLP (Multilayer Perceptron) and NARNet (Nonlinear Autoregressive Network) are employed to predict the WQI, and classification models: SVM (Support Vector Machine), XGBoost (eXtreme Gradient Boosting), Decision Trees, and Random Forest are employed to predict the WQC.	The ensemble regression model shows promise in predicting water quality time series data but has limitations. Reliance on specific regressors may limit adaptability to different datasets, and parameter choices could affect its performance. Its generalizability to diverse datasets beyond the specific timeframe and region of provided observations is uncertain, necessitating further validation for broader applicability in water quality prediction.
A Divide-and-Conquer Method Based Ensemble Regression Model for Water Quality Prediction [2013 Springer-Verlag Berlin Heidelberg]	• Xuan Zou, • Guoyin Wang, • Guanglei Gou • Hong Li	Support Vector Machine, RBF Neural Network and Grey Model	This paper proposes a novel ensemble regression model to predict time series data of water quality. The proposed model consists of multiple regressors and a classifier. The model transforms the original time series data into subsequences by sliding window and divides it into several parts according to the fitness of regressor so that each regressor has advantages in a specific part. The classifier decides which part the new data should belong to so that the model could divide the whole prediction problem into small parts and conquer it after computing on only one part.	Despite the promising results, the proposed ensemble regression model for predicting water quality time series data has several limitations. Firstly, its performanceis contingent on the qualityand representativeness of the training data, which may impact its generalization to new scenarios. The model's complexity, introduced by combining multiple regressors and a classifier,poses challenges in terms of longer training times, , and reduced interpretability.

CHAPTER – 3

SYSTEM REQUIREMENTS SPECIFICATION

3.1 Specific Requirement

3.1.1. Hardware Specification

- System: Intel i5 8th Gen Processor or higher
- Hard Disk: 80 GB
- RAM: 8 GB
- Memory: 15 GB

3.1.2. Software Specification

- Integrated Development Environment (IDE): Jupyter Notebook or JupyterLab
- Operating System: Windows 11
- Languages: Python, HTML, CSS
- Database: water_quality dataset

3.2 Functional Requirements

Functional requirements for a water quality prediction project outline the essential capabilities and behaviours that the system must exhibit to meet user needs and project goals. Here are key functional requirements:

1. Data Preprocessing:
    - Functionality to clean and preprocess raw data, including handling missing values, outliers, and noise.
    - Capability to perform data normalization and scaling to prepare data for modelling.
2. Model Training:
    - Support for various machine learning algorithms for regression and classification (e.g., SVM, MLP, linear regression, instance-based learning, LSTM, SVR, NAR Net).
    - Ability to train models on historical water quality data to predict Biochemical Oxygen Demand (BOD) and Water Quality Index (WQI).

3. Model Evaluation:
    - Tools to evaluate model performance using metrics such as accuracy, precision, recall, F1 score, RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2 (coefficient of determination).
    - Functionality to perform cross-validation and hyperparameter tuning to optimize model performance.
4. Visualization and Reporting:
    - Visualization tools to display historical water quality data, prediction results, and model performance metrics.
    - Generation of reports summarizing water quality status and model insights for stakeholders.
5. User Interface:
    - An intuitive user interface for data input, visualization, and interaction with the prediction model.
    - Accessibility features to accommodate various user needs.
6. Scalability and Integration:
    - Scalability to handle increasing amounts of data and incorporate additional sensors or data sources.
    - Integration with external systems and databases for data sharing and interoperability.
7. Security and Privacy:
    - Ensuring data security and privacy through encryption, access controls, and compliance with relevant regulations.
8. Documentation and Support:
    - Comprehensive documentation for system operation, data management, and troubleshooting.
    - Support services to assist users with technical issues and system maintenance.

3.4 Non-Functional Requirements

Non-functional requirements (NFRs) specify the criteria that judge the operation of a system, rather than specific behaviors or functions. For a water quality prediction project, non-functional requirements ensure the system's performance, usability, reliability, and other quality attributes. Here are key non-functional requirements:

1. Performance:
    - The system should process and analyze incoming data within a specified time frame to provide real-time or near-real-time predictions.
    - The model training and evaluation processes should be optimized for efficiency to handle large datasets without significant delays.

2. Scalability:
    - The system should be scalable to accommodate increasing amounts of data and additional sensors or data sources without compromising performance.
    - It should support horizontal and vertical scaling to handle growing user demands and data volumes.

3. Reliability:
    - The system must be reliable, with minimal downtime, ensuring continuous data collection, processing, and prediction.
    - It should include mechanisms for automatic recovery and fault tolerance in case of failures.

4. Availability:
    - The system should be available 24/7, especially if it is used for critical water quality monitoring and prediction.
    - High availability should be ensured through redundancy and failover mechanisms.

5. Security:
    - Data should be protected through encryption, both at rest and in transit, to prevent unauthorized access and tampering.
    - The system should implement robust access control measures to ensure that only authorized personnel can access sensitive data and functionalities.

6. Usability:
    - The user interface should be intuitive and user-friendly, allowing users to easily interact with the system, input data, and interpret results.
    - The system should provide clear visualizations and reports that are easy to understand for non-technical stakeholders.

7. Maintainability:
    - The system should be designed for easy maintenance, with modular components that can be updated or replaced without significant disruptions.
    - It should include comprehensive documentation to assist with maintenance and troubleshooting.

8. Interoperability:
    - The system should be able to integrate seamlessly with other systems, databases, and IoT devices, supporting standard data exchange formats and protocols.
    - It should facilitate data sharing and interoperability with external applications and stakeholders.

9. Efficiency:
    - The system should use computational and memory resources efficiently to minimize operational costs and environmental impact.
    - It should be optimized to reduce power consumption and maximize the use of available resources.

10. Accuracy:
    - The system should ensure high accuracy in data collection, processing, and prediction, minimizing errors and false positives/negatives.
    - It should include mechanisms for continuous validation and improvement of prediction models to maintain accuracy over time.

11. Responsiveness:
    - The system should respond promptly to user inputs and system events, providing timely feedback and updates.
    - It should ensure that users receive notifications and alerts without significant delays.

3.5 Performance Requirement

Performance requirements for a water quality prediction project define the expected levels of performance the system must achieve to be effective and efficient. Here are key performance requirements:

1. Response Time:
    - The system should provide predictions and visualizations within 5 seconds of receiving new data.
    - Alerts and notifications should be generated and sent within 2 seconds of detecting water quality issues.

2. Latency:
    - The end-to-end latency from data collection to prediction should not exceed 10 seconds.
    - Data transmission from sensors to the central processing unit should have a latency of less than 1 second.

3. Scalability:
    - The system should support scaling to accommodate up to 1,000 sensors and 10,000 data points per second.
    - It should maintain performance levels when scaling up to 10 times the initial workload.

4. Data Processing Speed:
    - The model training process should complete within 30 minutes for datasets up to 1GB in size.
    - Data preprocessing, including cleaning and normalization, should be completed within 10 seconds for each batch of 1,000 records.

5. Model Update Frequency:
    - The system should support updating predictive models at least once per day to incorporate new data.
    - Model retraining should not disrupt ongoing predictions and should run as a background process.

6. Resource Utilization:
    - CPU utilization should remain below 70% during peak processing times.
    - Memory utilization should not exceed 75% of available RAM to prevent swapping and ensure smooth operation.

7. Availability:
    - The system should achieve 99.9% uptime, ensuring high availability and minimal downtime.
    - Scheduled maintenance should not exceed 1 hour per month and should be performed during off-peak hours.

8. Accuracy:
    - The predictive model should achieve an accuracy of at least 95% in predicting water quality index (WQI).
    - The mean absolute error (MAE) of predictions should be less than 5% of the actual value.

9. Efficiency:
    - The system should optimize computational resources, ensuring energy-efficient operations.
    - It should leverage parallel processing and hardware acceleration (e.g., GPUs) to improve processing speed.

CHAPTER – 4
SYSTEM ANALYSIS

4.1 Existing System

The existing system is a project aimed at predicting water quality considering multiple attributes. It involves collecting data on parameters like pH, dissolved oxygen (DO), Biochemical Oxygen Demand (BOD), Total Suspended Solids (TSS), and Nitrate-Nitrogen (NO3-N).

Machine learning models, trained on historical data, classify the water as potable or not-potable based on the input parameters. Any one among the classification algorithm has used to classify the input data.

Existing systems preprocessing procedures are not much effective while removing outliers and handling the missing values. Also the accuracy achieved in existing systems was less.

4.1.1 Limitations

- No comparison has been considered among different ml models
- The accuracy achieved was less than 70%
- Most of the systems does not provide user interface

4.2 Proposed System

The proposed system for water quality prediction integrates advanced technologies to ensure accurate and efficient water resource management and environmental monitoring. It involves collecting data on parameters like pH, hardness, solids, chloramines, sulfates, conductivity, organic carbon, Trihalomethanes, Turbidity.

This data is undergoing preprocessing, including cleaning, normalization, and feature engineering. A robust database stores this data, which is then visualized through interactive dashboards and geospatial tools. Machine learning models, trained on historical data, classify the water as potable or not-potable based on the input parameters.

Various classification algorithms are considered to best fit the data. Continuous model retraining and system monitoring ensure the system adapts to changing conditions and maintains high performance. This integrated approach enhances the accuracy and efficiency of water quality assessments, providing a vital tool for sustainable water management.

4.2.1 Advantages

- Considered multiple classification algorithms
- Better user interface design
- Better training and testing accuracy
- Cross validation has used

CHAPTER – 5
SYSTEM DESIGN

5.1 Project Modules

A water quality prediction model that uses various attributes with classification algorithm can be divided into following modules:

Data Collection: In this module, a large dataset consisting different dimensions of water is collected. This dataset can be obtained from various sources such as online databases or by collecting data through surveys and experiments.

Pre-processing: This module involves pre-processing the collected data to remove noise, outliers, and inconsistencies. This step also involves data augmentation to increase the diversity of the dataset.

Model Training

Logistic Regression: Train a logistic regression model to classify the water quality data.

K-Nearest Neighbors (KNN): Train a KNN model to classify the water quality data based on the proximity of data points.

Decision Tree: Train a decision tree model to classify the water quality data using decision rules.

Random Forest: Train a random forest model, an ensemble of decision trees, to improve classification accuracy.

Support Vector Machine (SVM): Train an SVM model to classify the water quality data by finding the optimal hyperplane.

XGBoost: Train an XGBoost model, an optimized gradient boosting algorithm, to classify the water quality data.

Model Evaluation

Hypertuning: Perform hyperparameter tuning to optimize the parameters of the machine learning models for better performance.

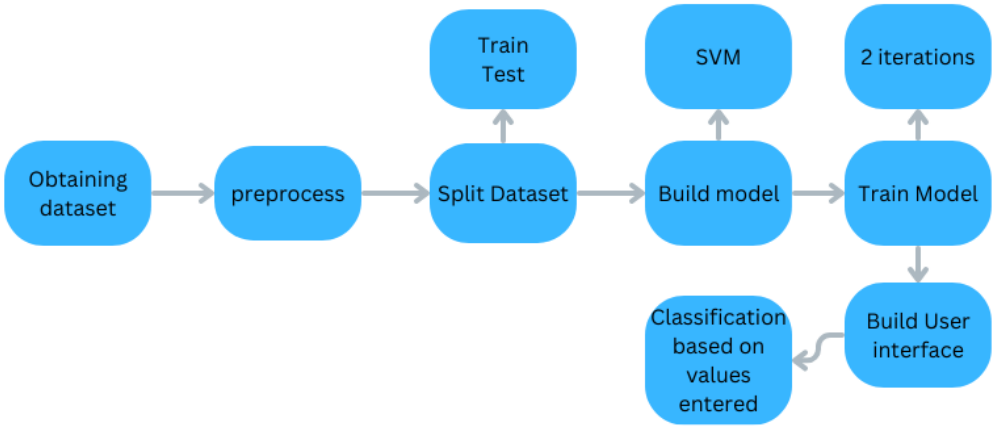
Algorithm Comparison: Compare the performance of different algorithms based on metrics such as accuracy, precision, recall, and F1-score.

Cross Validation: Use cross-validation techniques to evaluate the robustness and generalizability of the models.

Result and Interpretation:

Using SVM: The final results are interpreted using the SVM model, which is selected based on its performance during the evaluation phase. This model is used to classify the input dataset into potable or not-potable classes.

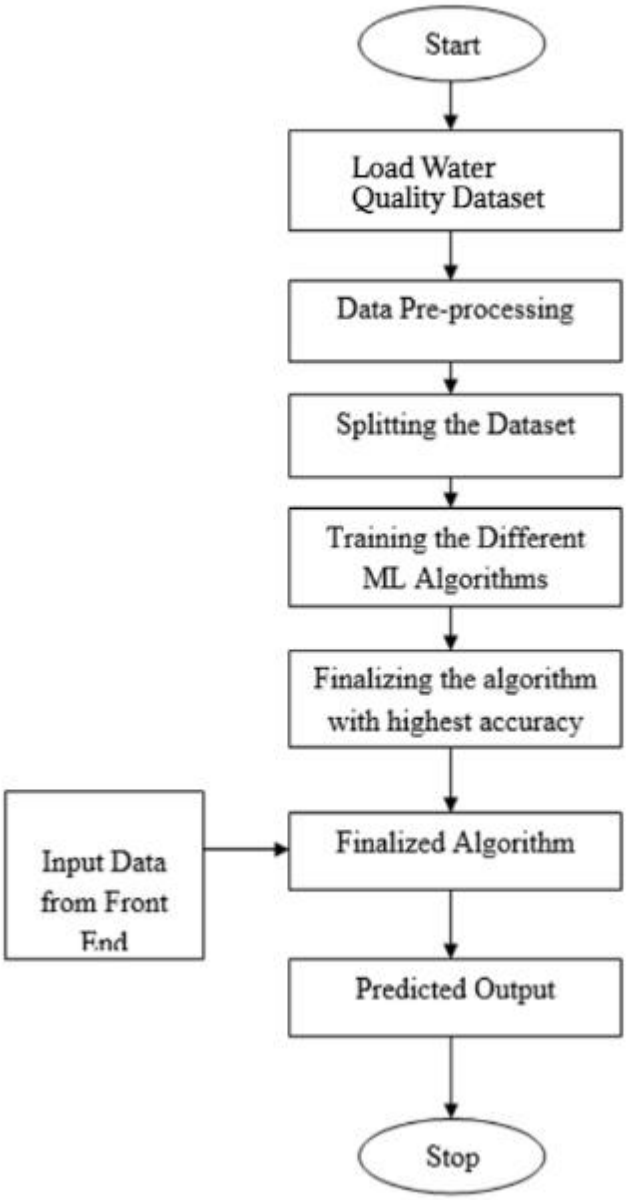
5.2 System Architecture



5.1 System Architecture of water quality analysis

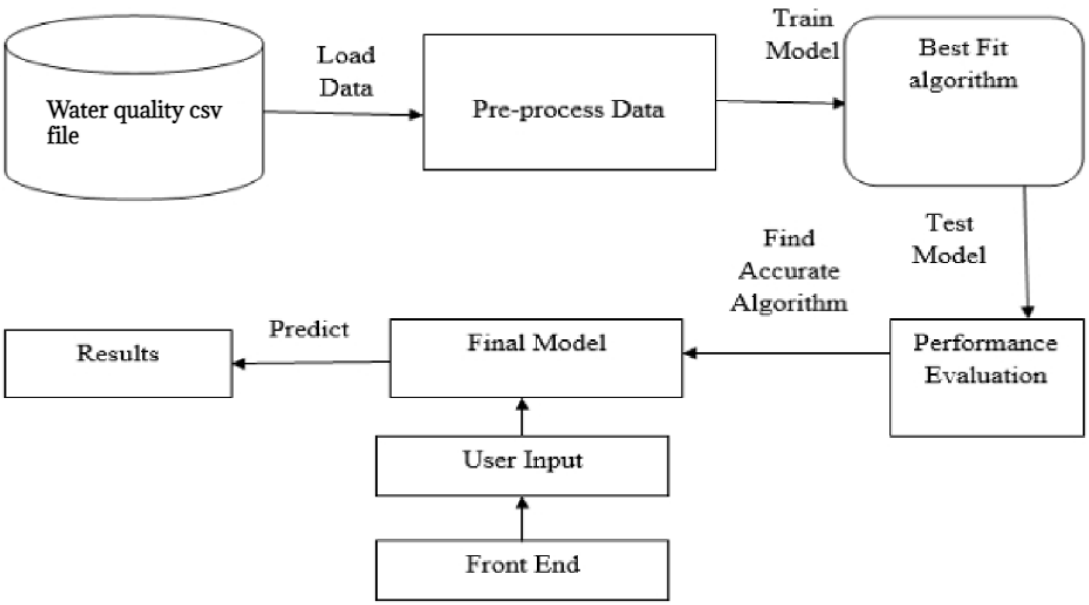
The system architecture for the water quality prediction model begins with obtaining a dataset containing key water quality parameters. The data undergoes preprocessing to clean and normalize it, ensuring its suitability for training. The preprocessed data is then split into training and testing sets. A Support Vector Machine (SVM) model is built and trained in two iterations to optimize its performance. After training, a user interface is developed, allowing users to input water quality parameters and receive classifications indicating whether the water is potable or not. This system ensures an accurate and efficient approach to water quality prediction, facilitating effective water resource management and environmental monitoring.

5.3 Control flow Diagram



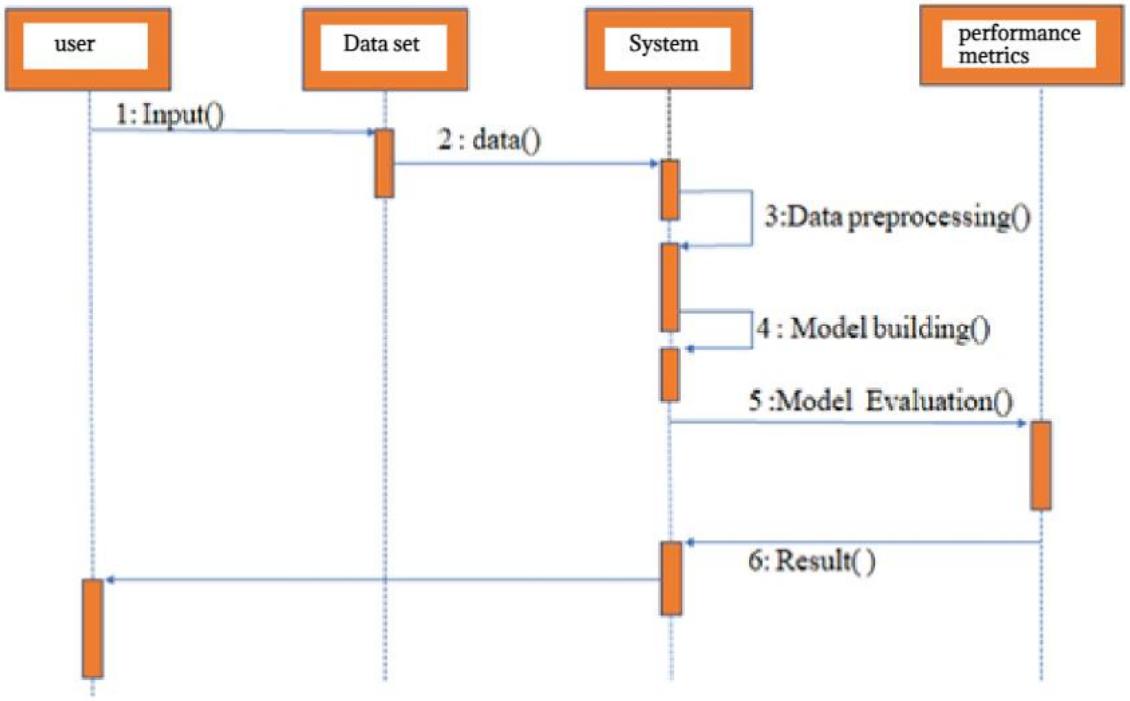
5.2 Control Flow diagram of water quality analysis

5.4 Data flow Diagram



5.3 Data Flow diagram of water quality analysis

5.5 Sequence Diagram



5.4 Sequence diagram of water quality analysis

CHAPTER – 6
IMPLEMENTATION

6.1 Concept

The approached used in this study involved four major stages. First stage is data processing where the initial analytics of the raw data is performed. This process includes data cleaning, transforming and source data correction. and detection of inconsistency within the records. The raw data is imported and managed in Python 3.7 where the data exploration analysis (EDA) started.

The goal of EDA was to discover useful information and preparing the technically correct data for preprocessing to become the consistent data that will be used for the rest of the modeling. Once the data is considered consistent data, preprocessing and modeling was conducted using SKlearn package where the dataset is split in a train-test split of 80% to training and 20% to testing. The dataset is then scaled so that the data points fit within an appropriate scale so that higher value ranges do not dominate when data point distances are calculated. The dataset is now ready for modeling.

Six algorithms were used: logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier, support vector modeling, and XGBoost algorithm. Two iterations of modeling were conducted, the first with default parameters and the second after hyper-tuning the parameters. After evaluation the models and their evaluation metrics, we further conducted cross validation on the top 3 performing models. Finally, some post-prediction analyses are discussed to finalize our results and review any recommendations with our final conclusions with the intent to deploy a robust water quality classifier.

6.2 Algorithm

Logistic Regression: Logistic regression is a statistical method similar to linear regression that estimates the probability of a binary event occurring. Based on the given independent variables in a dataset the model predicts the probability a classification will occur using log odd ratios and an iterative maximum likelihood method (Hoffman, 2019).

K-Nearest Neighbour Regression: The k-nearest neighbor algorithm assumes that similar things exist in close proximity. It is a non-parametric classifier which uses similar data points that are close to each other to make classifications. By storing all available cases, it will classify new cases based on similarity measures by pattern recognition.

Decision Tree Classifier: A decision tree classifier is a flowchart-like structure where attributes or variables of the dataset are represented as nodes and the branches from each node represent a decision rule to another node. Eventually, a decision will be reach to a final leaf-node that will be either one outcome or the other, in this case, potable or non-potable.

Random Forest Classifier: Random Forest classifier consists of a large number of individual decision tress that operate as a group or a 'forest'. Each individual tree will classify and return a prediction and the prediction with the most votes will become the final prediction. The fundamental concept of random forest is majority wins and is beneficial in datasets where the attributes or predictor variables have low correlation. The idea of having multiple decision trees essentially protects the forest from individual error (Yiu, 2019).

Support Vector Machine Classifier: Support Vector Machine (SVM) is a linear model for classification and creates a line or a hyperplane which separates the data into classes. The hyperplane in an n-dimensional Euclidean space is flat, n-1 dimensional subset of that space that divides it into two disconnected parts (Pupale, 2018).

XGBoost Classifier: XGBoost stands for extreme gradient boosting, and is an implementation of gradient boosted decision trees designed for speed and performance. Gradient boosting is a technique where new models are created to correct the errors made by previous models. The results are combined to make the final prediction (Brownlee, 2016).

6.3 Functional Modules

6.3.1 Obtain a Dataset from Kaggle:

The first step in building the website is to obtain dataset. A large dataset consisting different dimensions of water is collected. This dataset can be obtained from various sources such as online databases or by collecting data through surveys and experiments.

6.3.2 Pre-process the Dataset:

This module involves pre-processing the collected data to remove noise, outliers, and inconsistencies. This step also involves data augmentation to increase the diversity of the dataset.

6.3.3 Exploratory Data analysis:

Plotting the distribution of the data within each of the predictor variables show both non potable and potable records displaying a normal/Gaussian distribution pattern. Using box-plot, we can see the data distribution of the nine predictor variables and clearly identify outlier data points outside minimum and maximum whiskers within all variables. Using a pair-plot to also visualize the relationships between the predictor variables, the pairwise relationships did not yield any obvious linear relationships.

6.3.4 Build Machine Learning Model:

The next step is to build a machine learning model that can classify the water in the dataset. Various popular classification models were used while building.

6.3.5 Train the model:

Once the model has been built, it needs to be trained on the pre-processed and augmented dataset. This involves splitting the dataset into training and validation sets, and then feeding the images into the model in batches.

6.3.6 Test the model:

After training the model, it needs to be tested on a separate test set to evaluate its performance. This involves feeding the data rows into the model and comparing the predicted emotions to the potability column. Metrics such as accuracy, precision, recall, and F1 score can be used to evaluate the performance of the model.

6.3.7 Evaluate the Model:

Once the model has been tested, it needs to be evaluated to determine if it is suitable for the water prediction task. This involves analysing the model's strengths and weaknesses, as well as its performance on different emotions.

6.3.8 Build HTML for Frontend:

With the model in place, the next step is to build the HTML for the frontend of the website. This involves designing a user interface that allows users to interact with the website and select the emotion that they want to listen to music for.

6.3.9 Use Flask for Backend:

The backend of the website can be built using Flask, a lightweight web application framework for Python. Flask can be used to handle user requests, call the deep learning model to predict emotions, and serve up the appropriate music recommendations.

CHAPTER – 7
TESTING

7.1 Methods of Testing

Software testing is a process, to evaluate the functionality of a software application with an intent to find whether the developed software met the specified requirements or not and to identify the defects to ensure that the product is defect free in order to produce the quality product.

7.1.1 Unit Testing: This type of testing isolates and verifies the functionality of individual code components, like data processing or training routines. This ensures each part works as intended before integration, catching errors early on and guaranteeing proper data handling, model training, and prediction behavior.

7.1.2 System Testing: This type of testing verifies the complete machine learning system, ensuring it produces the intended results when fed specific data. This involves testing functionalities like training, inference (making predictions), and overall system performance. System testing goes beyond individual components and delves into how they interact as a whole. System testing simulates real-world use cases, feeding the system various data combinations to assess its ability to handle diverse scenarios. This can uncover unexpected issues, like the system struggling with specific data formats or producing nonsensical outputs under certain conditions.

7.1.3 Functional Testing: This type of testing is concerned with ensuring that the system works correctly based on the requirements and specifications. In case of functional testing for a water quality prediction system would involve checking if it delivers accurate forecasts based on the historical data.

7.1.4 Integration Testing: This type of testing involves testing how different modules or components of the system work together to ensure that they are compatible and that the overall system functions correctly. In case of water quality prediction system this goes beyond individual components and ensures they work together seamlessly. This involves verifying the smooth exchange of data between different parts. For instance, are real-time or historical data (temperature, pH) successfully transferred from the data acquisition module to the prediction model. Finally, does the visualization component flawlessly display these predictions (graphs, charts) for clear user understanding. By ensuring these pieces work together as a whole, integration testing guarantees a cohesive system that delivers accurate and comprehensive water quality predictions.

7.1.5 User Acceptance Testing: This type of testing ensures that the system meets the expectations of the stakeholders and end-users. In case of water quality prediction, it ensures system meets user needs. Here, predictions should align with user expectations of real-world water quality. This involves understanding those expectations. Users expect the system's forecasts, based on chosen settings and data, to accurately reflect actual water quality. Acceptance testing validates this by comparing predictions with verified measurements from reliable sources.

Overall, a comprehensive testing strategy should involve a combination of these types of testing to ensure that the water quality prediction system is functioning correctly, meets the user's expectations, and is secure and reliable.

7.2 Test Cases

Test ID	Input Data	Expected Output	Actual Output	Status
TC_1	Testing training and testing data	All matching outputs	All matching outputs	Pass
TC_2	Testing accuracy	80-85%	87%	Pass
TC_3	Real time data processing	Should run smoothly without error	Running smoothly without error	Pass
TC_4	Changing inputs to see change in prediction	Result should be changed	Result Changed	Pass

Table 7.1 Test Cases

CHAPTER – 8
PERFORMANCE ANALYSIS

Performance analysis is the process of evaluating the efficiency and effectiveness of a system or application in meeting its intended objectives. It involves measuring and analyzing various performance factors, such as response time, throughput, accuracy, resource utilization, and robustness. The purpose of performance analysis is to identify performance issues, bottlenecks, and areas for improvement, and to optimize the system or application to improve its overall performance.

Performance analysis is an important aspect of software development, as it helps ensure that the system or application can handle a high volume of requests and provide fast, accurate, and satisfying results to users. It also helps identify areas where the system or application can be optimized to use computing resources more efficiently and effectively, reducing costs and improving scalability.

Performance analysis can be performed using a variety of tools and techniques, such as load testing, stress testing, A/B testing, profiling, and monitoring. These tools provide quantitative data on the performance of the system or application, which can be used to identify performance issues and to track the effectiveness of performance optimizations over time.

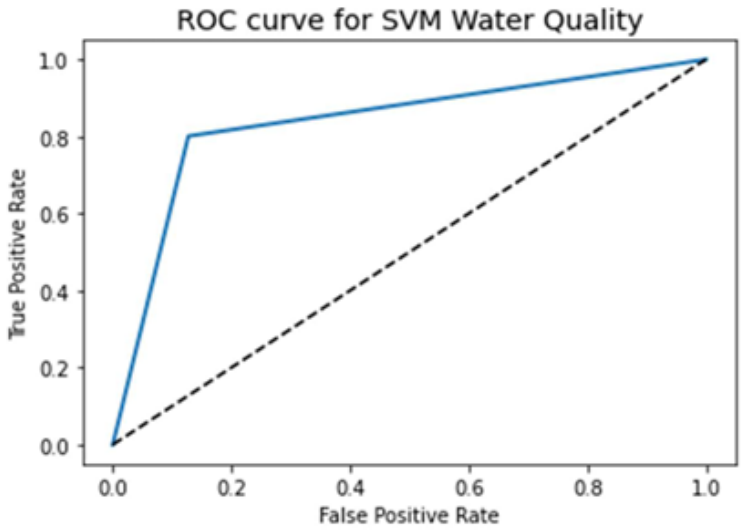


Figure 8.1 ROC Curve of SVM Model after second iteration.

Here are some factors that we considered when performing a performance analysis of our predictive water analysis system:

Accuracy: It measures the overall correctness of your model's predictions and is calculated as the number of correct predictions divided by the total number of predictions.

Precision: It measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. Precision is useful when the cost of false positives is high.

Recall (Sensitivity or True Positive Rate): It measures the proportion of correctly predicted positive instances out of the actual positive instances. Recall is useful when the cost of false negatives is high.

	precision	recall	f1-score	support
0	0.87	0.74	0.80	400
1	0.76	0.88	0.82	372
accuracy			0.81	772
macro avg	0.82	0.81	0.81	772
weighted avg	0.82	0.81	0.81	772

Table 8.1 Performance Analysis

F1 Score: It is the harmonic mean of precision and recall and provides a balanced measure of both metrics. F1 score is often used when you want to find a balance between precision and recall.

Specificity (True Negative Rate): It measures the proportion of correctly predicted negative instances out of the actual negative instances.

Response Time: Response time is the time it takes for the system to provide a recommendation after the user gives the input. A predictive water analysis system should be able to provide portability of water in real-time, without any lag or delay

Throughput: Throughput of a water quality analysis system refers to the amount of water samples that the system can analyze within a given time period. It is a measure of the system's efficiency and capacity to process and provide results for water quality tests. Higher throughput means that the system can handle a larger volume of samples quickly, which is crucial in environments where timely and frequent monitoring is required, such as in water treatment facilities, environmental monitoring stations, or industrial processes.

Scalability: Scalability of a water quality analysis system refers to the system's ability to efficiently handle an increasing volume of work or its potential to be enlarged to accommodate growth. Therefore, scalability in a water quality analysis system ensures that it remains effective, efficient, and reliable as demands and conditions evolve over time.

Resource Utilization: Resource Utilization in the context of a water quality analysis system refers to how efficiently and effectively the system uses its available resources to perform water quality testing and analysis. Effective resource utilization in a water quality analysis system ensures that the system operates smoothly, cost-effectively, and sustainably, while delivering high-quality and accurate water quality data. This is crucial for maintaining safe and healthy water standards and for making informed decisions related to water management and policy.

Robustness: The robustness of a water quality analysis system refers to its ability to consistently produce reliable, accurate, and precise results under a variety of conditions. This includes the system's capacity to handle different types of water samples, resist interference from potential contaminants, and function effectively despite environmental variations or operational challenges.

CHAPTER – 9

CONCLUSION AND FUTURE ENHANCEMENT

Conclusion

- We have built a Support Vector Machine (SVM) ML model that tackles the critical issue of water safety and demonstrates the power of machine learning in solving real-world problems.
- This model gives valuable insights into the complexity of assessing water quality.
- This model was trained using 2000+ data collected from various sources.
- In this analysis there is exploration, analysis, and implementation of various models.
- The project underscores the potential of data-driven solutions in safeguarding water.

Future Enhancement

- Continuous refinement and optimization of the SVM model can enhance its predictive capabilities.
- Fine-tuning hyperparameters and exploring advanced SVM configurations may lead to improved accuracy.
- Integrating educational components into the application can empower users with knowledge about water quality indicators and the significance of the model predictions. These fosters informed decision-making regarding water consumption.
- Collaborating with water authorities and environmental agencies can contribute to a more comprehensive understanding of water quality dynamics. This collaboration can provide access to diverse datasets and domain expertise.

BIBLIOGRAPHY

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

2. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

3. Journal of Emerging Technologies and Innovative Research [ISSN: 2349-5162] jetir.org

4. Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag. https://doi.org/10.1007/978-1-4757-2440-0

5. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536. https://doi.org/10.1038/323533a0

6. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106. https://doi.org/10.1023/A:1022643204877

7. Haykin, S. (1998). Neural Networks: A Comprehensive Foundation. Prentice Hall.

8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html

9. Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. https://doi.org/10.1007/b98835

10. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

APPENDIX

Appendix A: Screen Shots

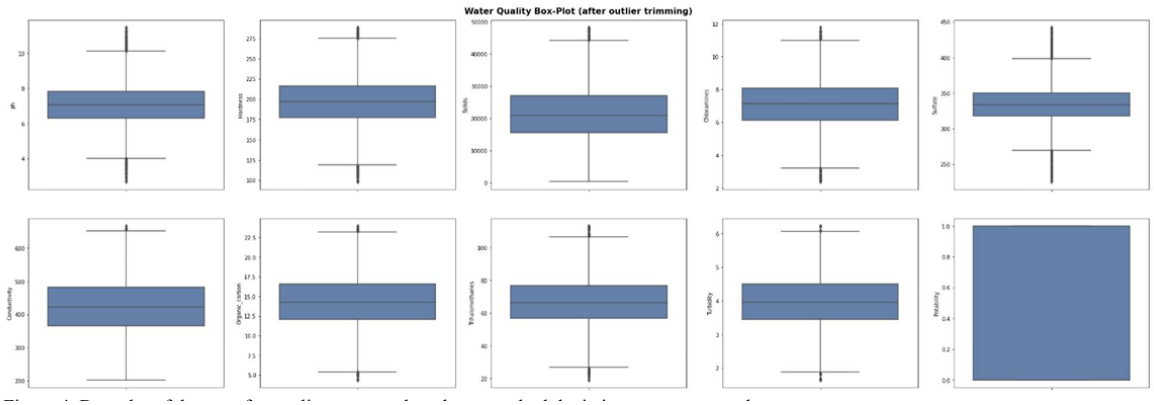


Fig A: Box-plot of dataset after outliers greater than three standard deviations were removed.

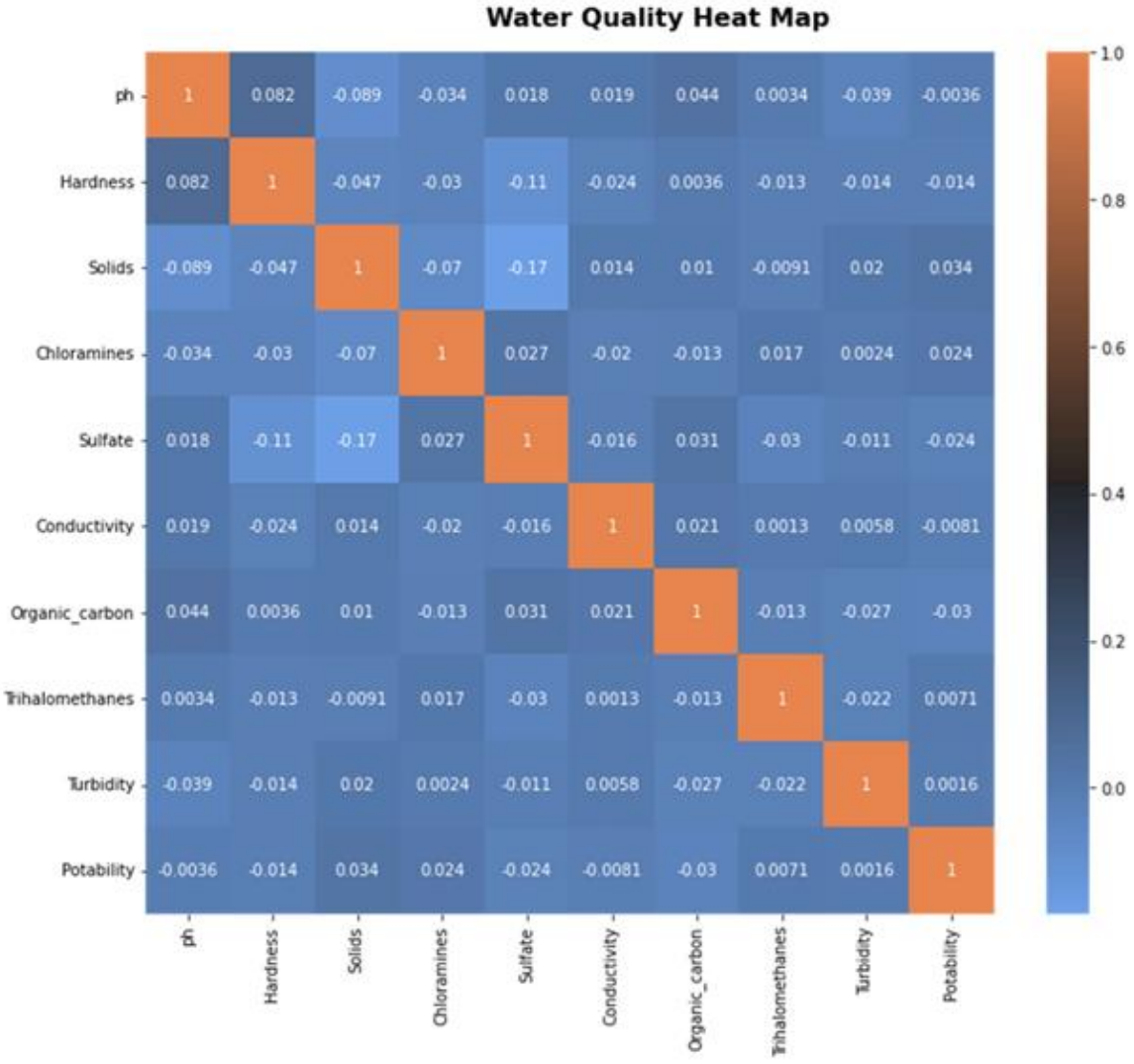


Fig B: Heatmap displaying correlation values.

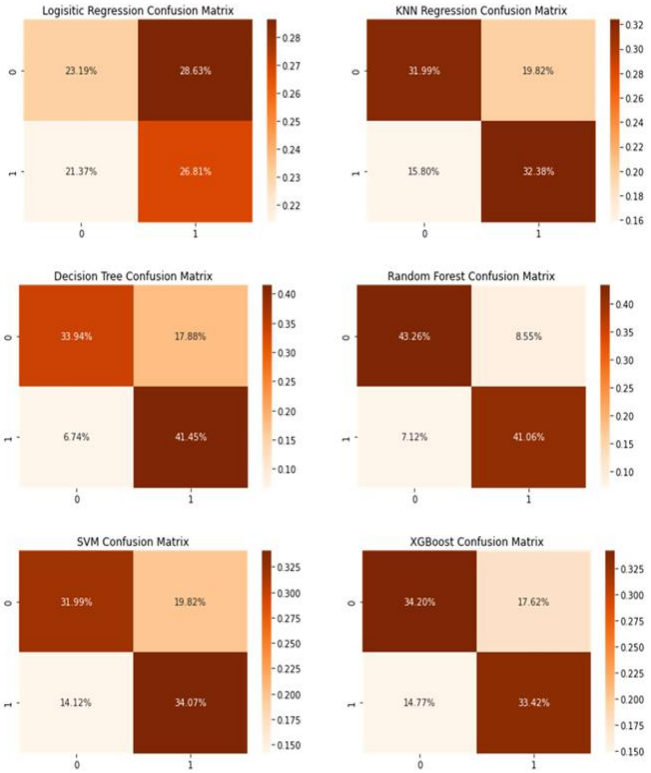


Fig C: Confusion Matrix of each algorithm after the first iteration of modeling

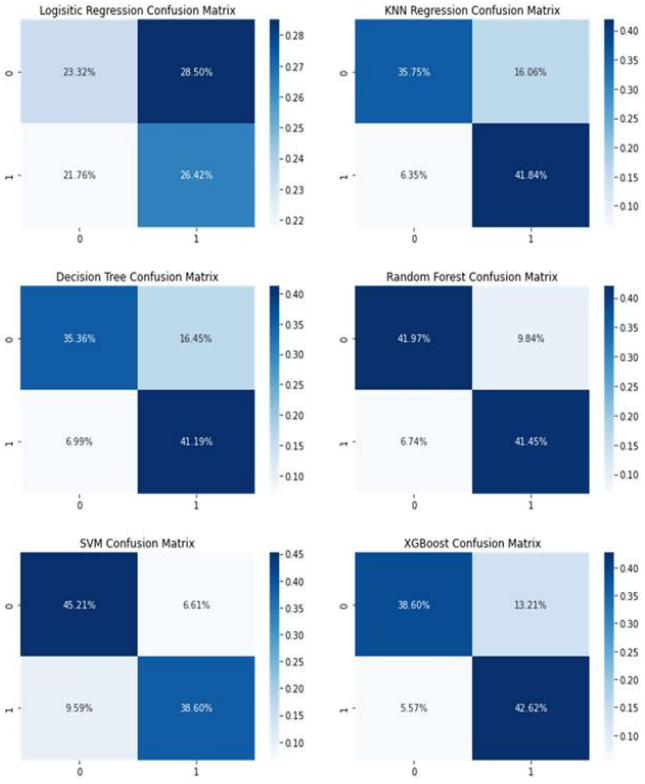


Fig D: Confusion Matrix of each algorithm after the second iteration of modeling

	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.843264	0.827676	0.852151	0.839735
2	Decision Tree	0.753886	0.698690	0.860215	0.771084
5	XGBoost	0.676166	0.654822	0.693548	0.673629
4	Support Vector	0.660622	0.632212	0.706989	0.667513
1	KNN Regression	0.643782	0.620347	0.672043	0.645161
0	Logistic Regression	0.500000	0.483645	0.556452	0.517500

Table 1: Evaluation metrics from the first iteration of modeling.

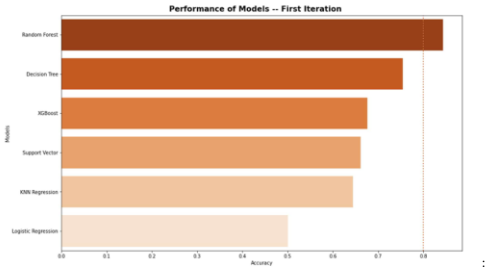


Fig E: Performance of each algorithm after the first iteration of modeling by accuracy.

	Model	Accuracy	Precision	Recall	F1 Score
4	Support Vector	0.838083	0.853868	0.801075	0.826630
3	Random Forest	0.834197	0.808081	0.860215	0.833333
5	XGBoost	0.812176	0.763341	0.884409	0.819427
1	KNN Regression	0.775907	0.722595	0.868280	0.788767
2	Decision Tree	0.765544	0.714607	0.854839	0.778458
0	Logistic Regression	0.497409	0.481132	0.548387	0.512563

Table 2: Evaluation metrics from the Second iteration of modeling.

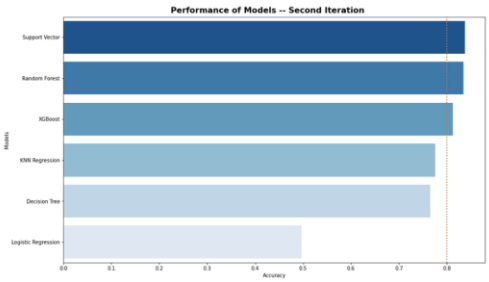


Fig F: Performance of each algorithm after the second iteration of modeling by accuracy.

Algorithm	Mean Accuracy Score	Standard Deviation
Random Forest	85.28 %	1.84 %
SVM	87.98 %	1.91 %
XGBoost	80.73 %	1.77%

Table 3: Results of K-Fold Cross Validation

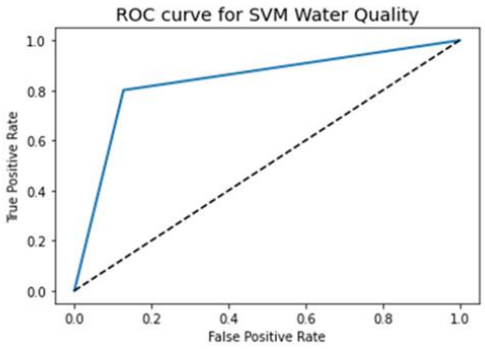


Fig G: ROC Curve of SVM Model after second iteration.

Appendix B: Abbreviations

ML: Machine Learning

DO: Dissolved Oxygen

BOD: Biochemical Oxygen Demand

TSS: Total Suspend Solids

NO3-N: Nitrate- Nitrogen

CSV: Comma Separated Values

KDE: Kernel Density Estimation

KNN: K-Nearest Neighbors

CNN: Convolutional Neural Network

HTML: Hyper Text Mark-up Language

CSS: Cascading Style Sheets

WQI: Water Quality Index

REPT: Reduced Error Pruning Tree

SVM: Support Vector Machine

PAPER PUBLICATION DETAILS

Published Paper

We published our paper on www.jetir.org, which is an International Open Access Journal.

JETIR is an UGC approved journal.

© 20XXJETIRMonth201X, Volume X, Issue Xwww.jetir.org (ISSN-2349-5162)

JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predictive Modelling of Urban Water Quality

Using Machine Learning

[1]Dr. Suma Swamy,[2] Siddhi Narayan,[3]Dharmitha V, [4]Y. Yasaswini

[1]Professor, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India

[2,3,4]B.E Student, Department of CSE, Sir M Visvesvaraya Institute of Technology, VTU, Bengaluru, India

Abstract: Urban water bodies face ongoing threats from pollution, necessitating advanced monitoring and prediction strategies to ensure timely intervention and resource preservation. This study explores the application of machine learning techniques for predictive modeling of water quality parameters in urban environments. Leveraging sensor data, satellite imagery, and historical records, the proposed models aim to forecast crucial indicators such as pH levels, dissolved oxygen, turbidity, and pollutant concentrations. The research emphasizes the development of accurate and robust machine learning algorithms capable of real-time monitoring, enabling early detection of potential contamination events. By integrating predictive analytics into water quality management systems, this approach facilitates proactive decision-making, contributing to the sustainable preservation of urban water resources. The findings hold significant implications for environmental monitoring, policy formulation, and the creation of smart city infrastructures aimed at safeguarding water quality in the face of growing urbanization and environmental challenges.

IndexTerms **- Urban water bodies, pollution, machine learning, predictive modelling, sensor data, satellite imagery, real-time monitoring, water quality management, smart city, environmental challenges.**

I. INTRODUCTION:

Water quality is a critical aspect of environmental health and human well-being, as it directly impacts ecosystems, agriculture, and human consumption. With the increasing pressures on water resources due to population growth, industrialization, and climate change, ensuring water quality has become a paramount concern. Traditional methods of water quality monitoring involve manual sampling and laboratory analysis, which are often time-consuming and costly.

In recent years, there has been a growing interest in leveraging machine learning (ML) techniques to predict and monitor water quality more efficiently. ML algorithms can analyse large datasets containing diverse water quality parameters, environmental variables, and historical records to make accurate predictions about water quality conditions. This approach offers the advantage of real-time or near-real-time monitoring, allowing for timely interventions and proactive management of water resources.

The utilization of machine learning (ML) for water quality prediction offers a plethora of benefits that address the challenges inherent in traditional monitoring methods. Firstly, ML algorithms enable the early detection of contamination events, providing a crucial advantage in swiftly implementing preventive measures to safeguard public health and environmental integrity. Moreover, the predictive modeling capabilities of ML empower stakeholders to anticipate changes in water quality by analyzing historical data alongside relevant environmental variables. This not only aids in proactive management but also facilitates optimized resource allocation, ensuring that interventions are strategically deployed where they are most needed.

II. RESEARCH OBJECTIVE

Development of an Integrated Machine Learning Model: The primary goal is to engineer a sophisticated machine learning model capable of integrating and processing diverse environmental parameters such as pH, dissolved oxygen, biochemical oxygen demand, total suspended solids, and nitrate-nitrogen. This model will be designed to handle real-time data variations, aiming to deliver robust and accurate predictions essential for efficient water resource management.

Implementation of a Water Quality Classification System: The research will also focus on establishing a reliable classification system that can discern whether water is potable or not based on its environmental and chemical characteristics. This system will incorporate advanced data preprocessing techniques to enhance the accuracy of the classification, providing a dependable tool for environmental agencies and policymakers.

`

© 20XXJETIRMonth201X, Volume X, Issue Xwww.jetir.org (ISSN-2349-5162)

Development of Data Visualization Modules: Another critical objective involves creating intuitive data visualization modules to effectively communicate the complexities of water quality data. These visualizations will assist in making the data comprehensible to various stakeholders, thereby supporting improved decision-making and increasing public awareness about water quality challenges.

III. LITERATURE REVIEW

In 2021, Ali Omran Al-Sultani [1], studied and developed ensemble machine learning models to predict BOD values in the Euphrates River, Iraq, using feature extraction techniques like Genetic Algorithm and PCA. It focused on Rohri Canal, highlighting the need for broader geographical representation. The study's limitations include neglecting temporal/spatial variability and a static dataset. Real-time data integration is suggested for proactive water management in smart cities, showing promise for future research in water quality prediction.

In 2023, Sarang Karim [2], proposed an IoT and ML framework for water quality analysis in Rohri Canal, Pakistan, using MLP for regression and Random Forest for classification to predict WQI and WQC. Results favor MLP for regression and Random Forest for classification, showing better performance with smaller datasets. Limitations include a short two-year dataset and the absence of a climate change variable. Future research should consider broader datasets, additional metrics, and advanced models for enhanced analysis.

In 2022, Bilal Aslam [3], aimed to enhance Water Quality Index predictions using machine learning on well water samples from North Pakistan. Hybrid algorithms outperformed standalone ones. Future work could explore algorithm performance over extended periods, include important parameters like COD and BOD over multiple years, consider deep learning algorithms like CNNs, conduct PCA tests, and analyse additional water quality variables for a robust study.

In 2023, Mahmoud Y Shams [4] study, focused on water quality prediction using machine learning, optimizing models like RF, Xgboost, GB, etc., through grid search. GB achieves 99.50% accuracy in classification (WQC), and MLP excels with a 99.8% R2 value in regression (WQI). Limitations include dataset specificity, potential parameter sensitivity oversight, limited assessment metrics, and absence of uncertainty analysis. Future research could explore broader datasets, additional metrics, and thorough sensitivity analysis for enhanced model performance evaluation.

In 2022, Mohamed Abbas [5], introduced AdCSO-sELM, enhancing ELM performance with dynamic parameter adjustment. Achieved 96.54% accuracy in classifying water potability with Kaggle dataset. Limitations include data dependency, dataset reliance, and generalizability issues. Novel dynamic adaptation may increase complexity and computational load. Performance may vary across environments, requiring extensive parameter tuning.

In 2020, Salwani Abdullah [6], inspired by the human body's remarkable waste filtration system, researchers have introduced a new optimization algorithm called Kidney-Inspired (KA). This algorithm mimics the way healthy kidneys work, iteratively evaluating potential solutions and discarding less favorable options, much like how kidneys filter waste products from the blood. The study demonstrates KA's potential in real-world applications like water quality prediction and cancer detection, suggesting its competency in handling complex tasks. However, the research also acknowledges limitations in a simplified version of KA known as Dual-KA. While Dual-KA streamlines the optimization process, it faces challenges. It may not fully capture the intricate biological processes of real kidneys, limiting its direct biological relevance. Additionally, Dual-KA might be more effective for specific types of problems, hindering its broader applicability. Furthermore, understanding how Dual-KA arrives at its decisions can be complex, and its ability to handle large-scale problems remains to be fully explored.

In 2013, Xuan Zou [7], proposed a new method for predicting water quality changes over time. It combines multiple prediction models (regressors) with a decision-maker (classifier). The data is divided into sections based on which model works best for that specific part. The classifier then directs new data to the most effective model, allowing the system to handle complex variations in water quality. While promising, the approach has limitations. The model's accuracy relies heavily on the quality and completeness of the training data. Additionally, the combination of multiple models increases training time and makes it more difficult to understand how the model arrives at its predictions.

In 2020, Bharat B. Gulyani [8], traditionally, measured Biochemical Oxygen Demand (BOD) in water is time-consuming and expensive. This research proposes a data-driven machine learning model for faster and cheaper BOD prediction. The model utilizes dimensionality reduction techniques to analyze only the most relevant data, improving efficiency. However, the authors acknowledge that the effectiveness of specific techniques might vary and call for further investigation into potential challenges like overfitting and outlier sensitivity. Additionally, they emphasize the need for more comprehensive evaluation metrics to assess the model's generalizability and real-world performance.

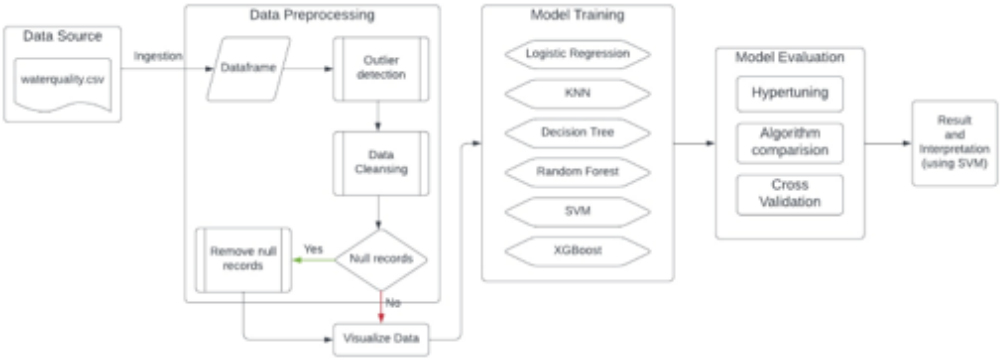
© 20XXJETIRMonth201X, Volume X, Issue Xwww.jetir.org (ISSN-2349-5162)

In 2009, Tianyou Cha [9], research introduced a new method using a mechanism model and hierarchical neural networks to predict water quality in wastewater treatment plants. This soft measurement approach aims to be more cost-effective than current online sensors. The hierarchical structure mimics the cascaded stages of the treatment process, potentially improving prediction accuracy within the reactors. While the study shows promise, the authors acknowledge limitations. Implementing the method could be complex, and its effectiveness relies on high-quality data. Additionally, adapting the model to real-time changes and different plant configurations might be challenging. Further research is needed to address these limitations before widespread adoption in wastewater treatment facilities.

In 2022, K. P. Rasheed [10], research introduced new models for predicting water quality in aquaculture settings. The models combine the strengths of Convolutional Neural Networks (CNNs) for capturing water quality characteristics and Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks for handling long-term trends in water data. These models achieved good results on test datasets. However, the study acknowledges limitations. The models might not adapt easily to different aquaculture systems due to variations in water quality factors. Additionally, they rely on specific settings (hyperparameters) that may need adjustments for new data or environments. Finally, while the most successful model (CNN-LSTM) performs well, its complex calculations could be challenging for real-time use in aquaculture management.

-

. IV. PROPOSED SYSTEM:



Data Preprocessing: It is a crucial step in preparing raw data for analysis and modeling. It involves creating a data frame from a CSV file and performing data cleansing, which includes handling outliers and null records.

- Outlier Detection: The data is scanned for outliers, which are data points that fall outside the expected range.
- Data Cleansing: This step involves removing or correcting any errors in the data, such as null values. Here, the process checks for null records and removes them if there are any.

Data Visualization: The distributions of each dimension in the dataset were visualized using histograms with KDE plots, while outliers were identified through box plots. Additionally, a heat map was employed to illustrate the correlations among all dimensions, providing a comprehensive overview of the data's relationships and anomalies.

Model Training: Various classification models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost and K-Nearest Neighbors, were trained on the water quality dataset. Their accuracies were evaluated, and confusion matrices were used to summarize the performance of each model, providing insights into their predictive accuracy and classification performance.

Model evaluation: It is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

- Hyper tuning: Training your model sequentially with different sets of hyperparameters.
- Cross-validation: Evaluating models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

Result: Post prediction analysis is discussed to finalize our results and review any recommendations with our final conclusions with intent to deploy a robust water quality classifier.

V. RESULTS AND CONCLUSIONS:

© 20XXJETIRMonth201X, Volume X, Issue Xwww.jetir.org (ISSN-2349-5162)

In the first iteration of modeling, the algorithms were all run just the default parameters of their respective functions. The confusion matrices of the algorithms are as follows:

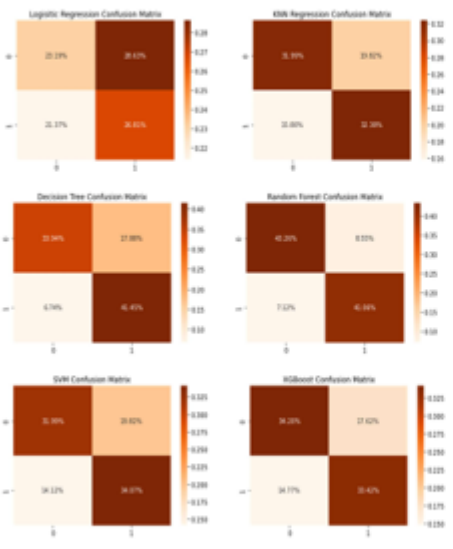


Fig A: Confusion Matrix of each algorithm after the first iteration of modeling

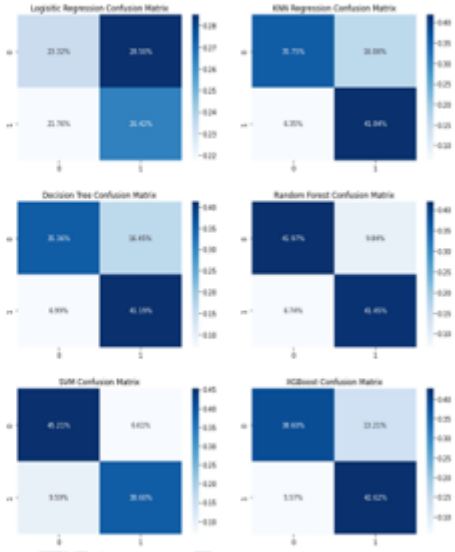


Fig B: Confusion Matrix of each algorithm after the second iteration of modeling

Statistically, type II errors would be considered more hazardous in particular for assessing water quality or potability. Returning a classification of a false positive would be detrimental to a community consuming unclean water. In reviewing the confusion matrices, the highest type II error occurs in Logistic Regression at 21.37% while the lowest type II error occurs in the Decision Tree algorithm at 7.12%.

	Model	Accuracy	Precision	Recall	F1 Score
3	Random Forest	0.843264	0.827676	0.852151	0.839735
2	Decision Tree	0.753886	0.698690	0.860215	0.771084
5	XGBoost	0.676166	0.654822	0.693548	0.673629
4	Support Vector	0.660622	0.632212	0.706989	0.667513
1	KNN Regression	0.643782	0.620347	0.672043	0.645161
0	Logistic Regression	0.500000	0.483645	0.556452	0.517500

Table: Evaluation metrics from the first iteration of modeling.

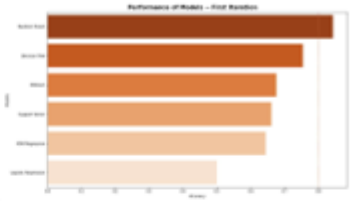


Fig C: Performance of each algorithm after the first iteration of modeling by accuracy.

After reviewing the evaluation metrics of each algorithm for first iteration of modeling, we observed that Random Forest algorithm performed the best with an accuracy of 84.33% while Logistic Regression performed the worst at 50.00%.

	Model	Accuracy	Precision	Recall	F1 Score
4	Support Vector	0.838083	0.853868	0.801075	0.826630
3	Random Forest	0.834197	0.808081	0.860215	0.833333
5	XGBoost	0.812176	0.763341	0.884409	0.819427
1	KNN Regression	0.775907	0.722595	0.868280	0.788767
2	Decision Tree	0.765544	0.714607	0.854839	0.778458
0	Logistic Regression	0.497409	0.481132	0.548387	0.512563

Table: Evaluation metrics from the Second iteration of modeling.

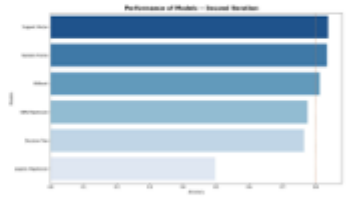


Fig D: Performance of each algorithm after the second iteration of modeling by accuracy.

Looking at the evaluation metrics after the second iteration we can see that hyper-tunning greatly increased the performance of nearly all the algorithms. SVM and Random Forest performed the best with the highest accuracy of 83.81% and 83.42% respectively. Logistic Regression remained the lowest performing algorithm with minute changes in overall performance.

© 20XXJETIRMonth201X, Volume X, Issue Xwww.jetir.org (ISSN-2349-5162)

Algorithm	Mean Accuracy Score	Standard Deviation
Random Forest	85.28 %	1.84 %
SVM	87.98 %	1.91 %
XGBoost	80.73 %	1.77%

Table 3: Results of K-Fold Cross Validation.

After the second iteration of model training, we selected the top three algorithms to apply cross validation to. Using the K-Fold Cross-Validation method, the consistent dataset (the dataset before train-test split) was used to be split into k number of subsets, where k-1 subsets are used to train the models and the last subset is kept for validation to test the models. The scores of each fold are then averaged to evaluate the overall performance of each model. Cross-validation using 10-folds, where 9 folds were used for training and 1 used for testing, returned higher accuracy results in all three algorithms: Random Forest, SVM, and XGBoost.

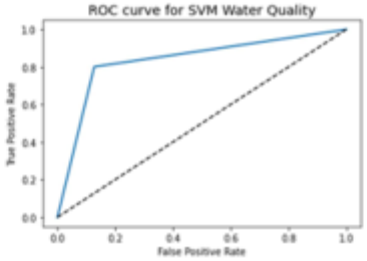


Fig 8: ROC Curve of SVM Model after second iteration.

Since SVM has the highest mean accuracy score after cross validation, we returned to the second iteration of model to produce a Receiver Operating Characteristic Curve (ROC Curve) graph to visualize the SVM model's performance with respects to their classification threshold levels. The ROC Curve plots the True Positive Rate (recall) against the False Positive Rate (type II error). We can also calculate the area under the curve (ROC AUC) which will allow us to understand the classifier's performance numerically as a perfect classifier is equal to 1.0. The ROC AUC for SVM after the second iteration was 0.8368 and the cross validated ROC AUC was 0.8674 which is consistent with the rest of our evaluation metrics.

VI. References

1. Proposition of New Ensemble Data-Intelligence Models for Surface Water Quality Prediction
[IEEE, Volume:9], 26 July 2021
2. Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality
[IEEE Volume 11], 14 September 2023
3. Water Quality Management Using Hybrid Machine Learning and Data Mining Algorithms: An Indexing Approach
[IEEE Volume 10], 10 November 2022
4. Water quality prediction using machine learning models based on grid search method
[Multimedia tools and Applications Journal], 29 September 2023
5. Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm
[IEEE Volume 11], 2022
6. Dual Kidney- Inspired Algorithm for Water Quality Prediction and Cancer Detection
[IEEE Volume 8], 2020
7. A Complete Proposed Framework for Coastal Water Quality Monitoring System With Algae Predictive Model
[IEEE Volume 9], 2021
8. Machine learning algorithms for efficient water quality prediction
[Modelling Earth Systems and Environments algorithm], 2021
9. Towards design of IoT and Machine Learning Embedded Frameworks for analysis and prediction of water quality

© 20XXJETIRMonth201X, Volume X, Issue Xwww.jetir.org (ISSN-2349-5162)

[IEEE Volume 11], 2023

10. A Divide-and-Conquer Method Based Ensemble Regression Model for Water Quality Prediction
2013, [Springer-Verlag Berlin Heidelberg]

11. Dimensionality Reduction for Water Quality Prediction from a Data Mining Perspective
2020, [Springer Nature Singapore Pte Ltd]

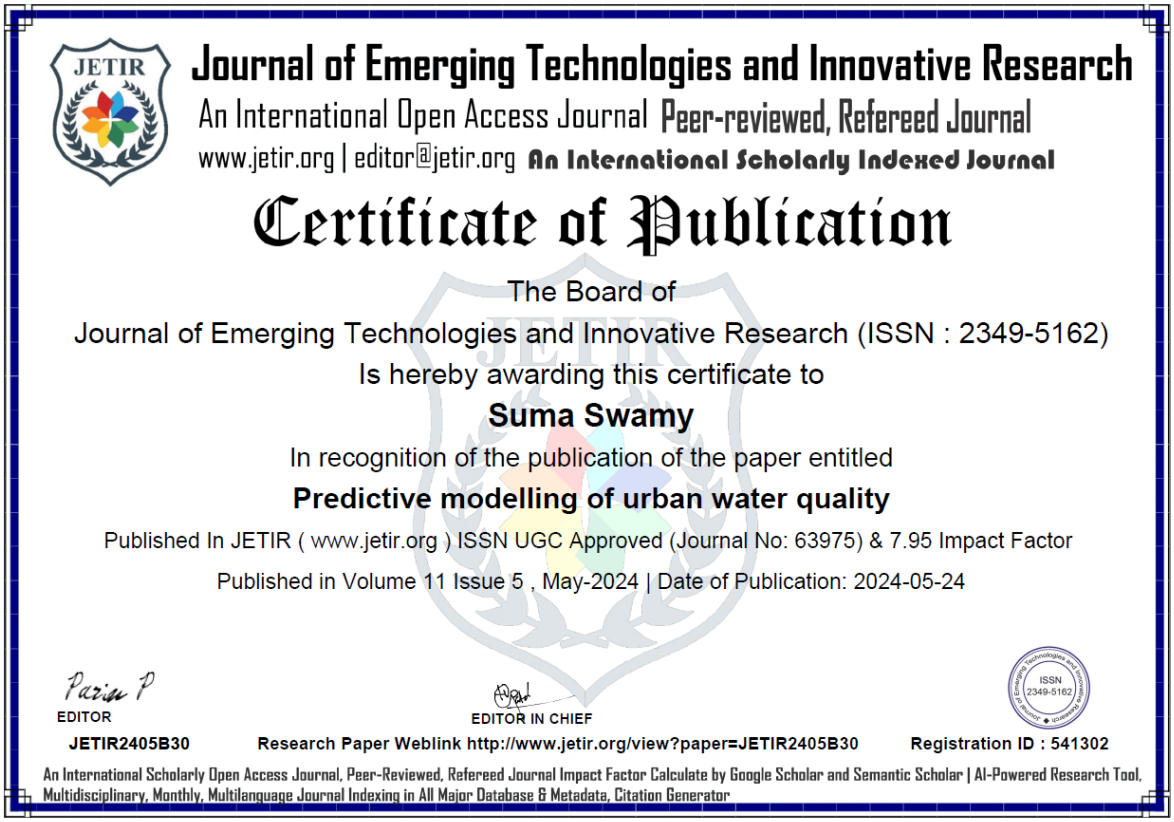
12. Hierarchical Neural Network Model for Water Quality Prediction in Wastewater Treatment Plants
2009, [Springer-Verlag Berlin Heidelberg]

13. Water Quality Prediction for Smart Aquaculture Using Hybrid Deep Learning Models
[IEEE Volume 10], 2022

14. Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation
2010, [Springer-Verlag London Limited]

Paper id | Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org

Certificates:



Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

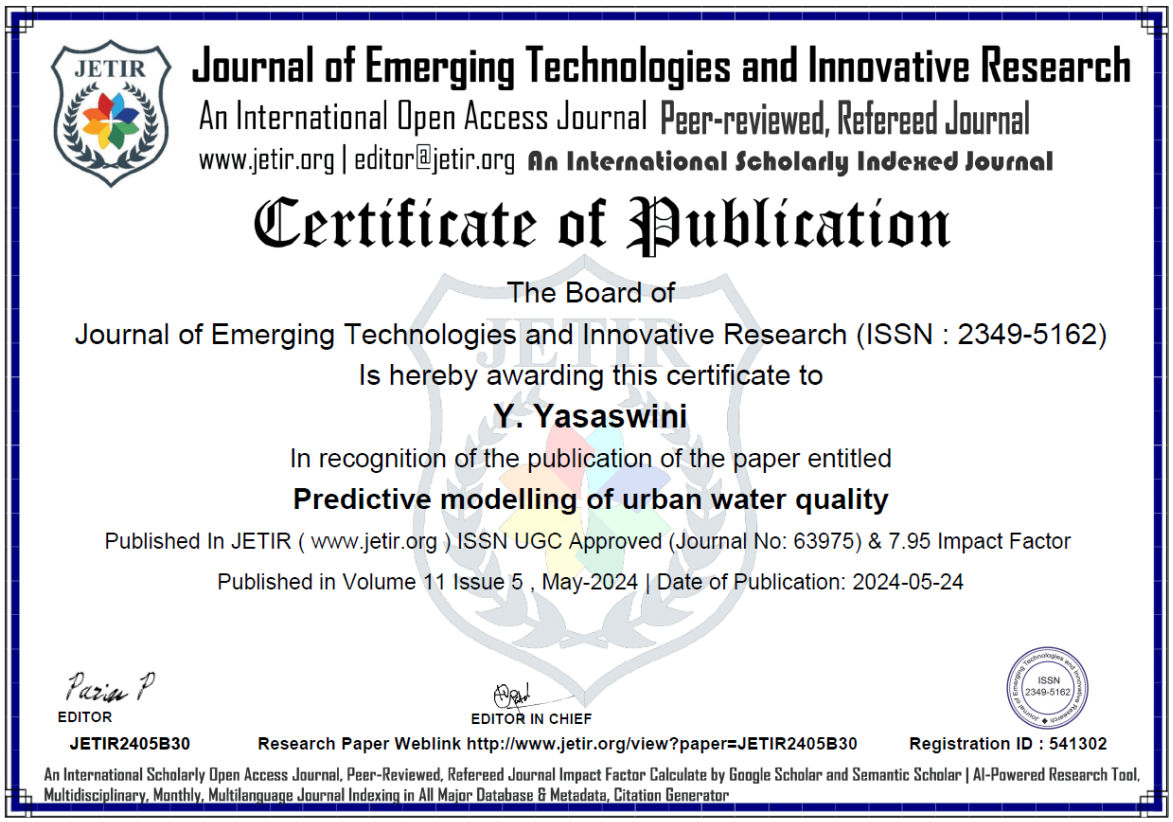
Certificate of Publication

The Board of
Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)
Is hereby awarding this certificate to
Suma Swamy
In recognition of the publication of the paper entitled
Predictive modelling of urban water quality
Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor
Published in Volume 11 Issue 5 , May-2024 | Date of Publication: 2024-05-24

EDITOR EDITOR IN CHIEF

JETIR2405B30 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2405B30 Registration ID : 541302

An International Scholarly Open Access Journal, Peer-Reviewed, Refereed Journal Impact Factor Calculate by Google Scholar and Semantic Scholar | AI-Powered Research Tool, Multidisciplinary, Monthly, Multilanguage Journal Indexing in All Major Database & Metadata, Citation Generator



Journal of Emerging Technologies and Innovative Research
An International Open Access Journal Peer-reviewed, Refereed Journal
www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

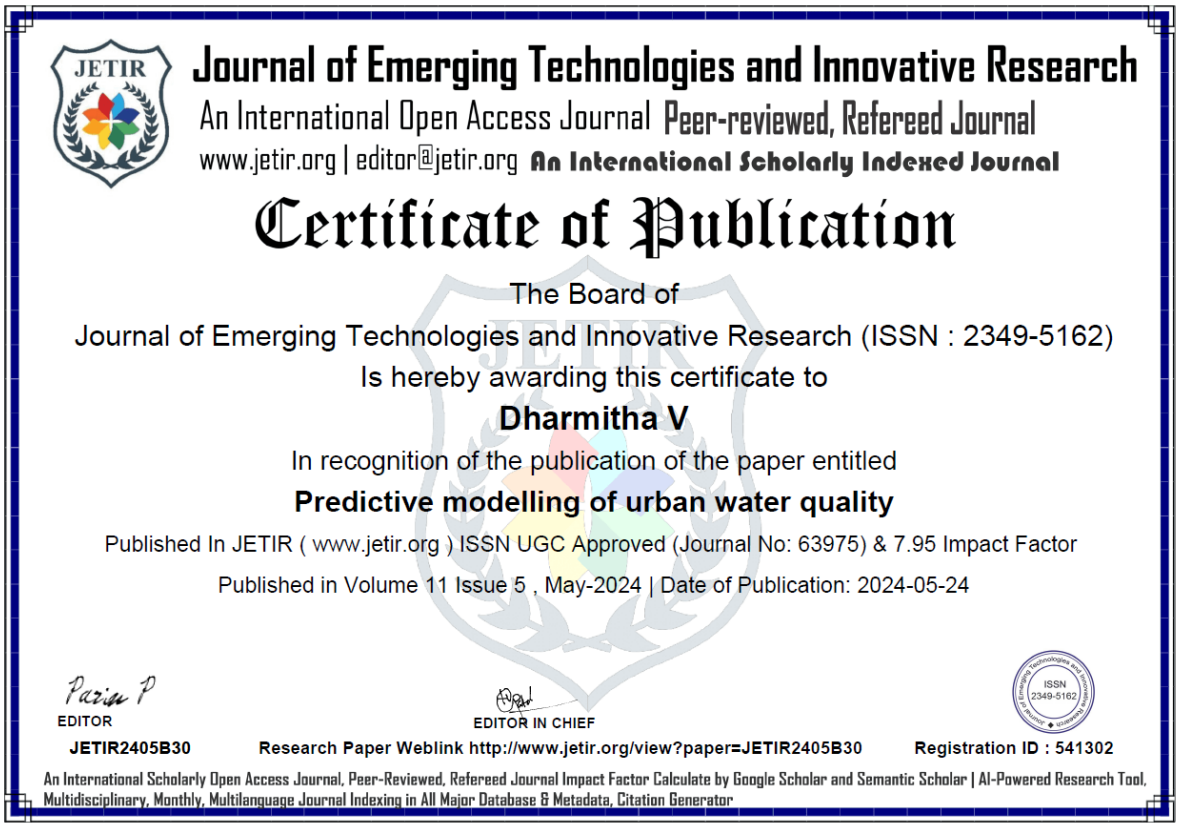
Certificate of Publication

The Board of
Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)
Is hereby awarding this certificate to
Y. Yasaswini
In recognition of the publication of the paper entitled
Predictive modelling of urban water quality
Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor
Published in Volume 11 Issue 5 , May-2024 | Date of Publication: 2024-05-24

EDITOR EDITOR IN CHIEF

JETIR2405B30 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2405B30 Registration ID : 541302

An International Scholarly Open Access Journal, Peer-Reviewed, Refereed Journal Impact Factor Calculate by Google Scholar and Semantic Scholar | AI-Powered Research Tool, Multidisciplinary, Monthly, Multilanguage Journal Indexing in All Major Database & Metadata, Citation Generator



Journal of Emerging Technologies and Innovative Research

An International Open Access Journal Peer-reviewed, Refereed Journal

www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

Dharmitha V

In recognition of the publication of the paper entitled

Predictive modelling of urban water quality

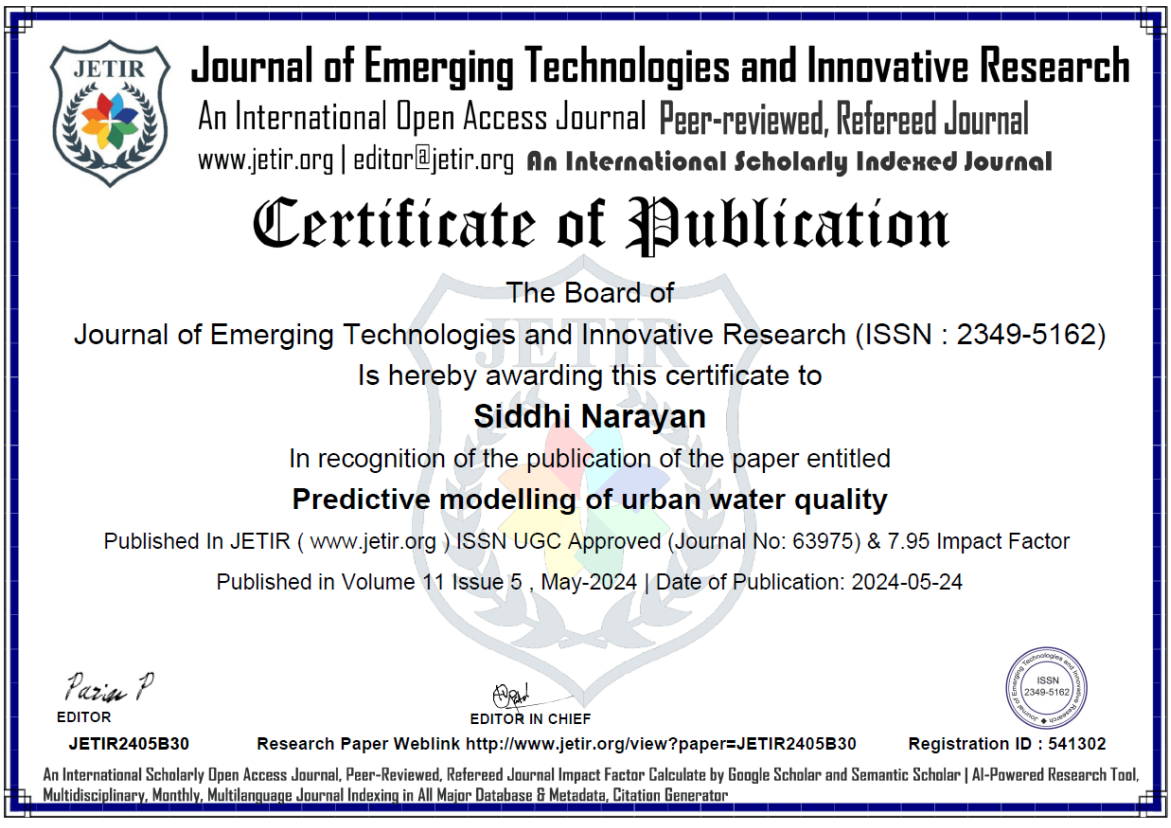
Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 11 Issue 5 , May-2024 | Date of Publication: 2024-05-24

EDITOR EDITOR IN CHIEF

JETIR2405B30 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2405B30 Registration ID : 541302

An International Scholarly Open Access Journal, Peer-Reviewed, Refereed Journal Impact Factor Calculate by Google Scholar and Semantic Scholar | AI-Powered Research Tool, Multidisciplinary, Monthly, Multilanguage Journal Indexing in All Major Database & Metadata, Citation Generator



Journal of Emerging Technologies and Innovative Research

An International Open Access Journal Peer-reviewed, Refereed Journal

www.jetir.org | editor@jetir.org An International Scholarly Indexed Journal

Certificate of Publication

The Board of

Journal of Emerging Technologies and Innovative Research (ISSN : 2349-5162)

Is hereby awarding this certificate to

Siddhi Narayan

In recognition of the publication of the paper entitled

Predictive modelling of urban water quality

Published In JETIR (www.jetir.org) ISSN UGC Approved (Journal No: 63975) & 7.95 Impact Factor

Published in Volume 11 Issue 5 , May-2024 | Date of Publication: 2024-05-24

EDITOR EDITOR IN CHIEF

JETIR2405B30 Research Paper Weblink http://www.jetir.org/view?paper=JETIR2405B30 Registration ID : 541302

An International Scholarly Open Access Journal, Peer-Reviewed, Refereed Journal Impact Factor Calculate by Google Scholar and Semantic Scholar | AI-Powered Research Tool, Multidisciplinary, Monthly, Multilanguage Journal Indexing in All Major Database & Metadata, Citation Generator