WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Water Quality Prediction and Classification Using Machine Learning

P. Firoz Khan*[1], Sk. Zohra Zaheen*[2], Dr. P. Shyam Sunder[3], Ms. K. Shirisha[4], Dr.Rajitha Kotoju[5], R. Mohan Krishna Ayyappa[6]

[1,2]Under Graduate Student, [3,4,5,6]Assistant Professor, [3,4]Mentors, [4,5]Coordinators
Department of Computer Science & Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India
pfirozkhan_cse2205b2@mgit.ac.in, szohrazaheen_cse2205c0@mgit.ac.in

ABSTRACT

Water quality assessment plays a vital role in environmental monitoring, public health, and sustainable water resource management. It helps in monitoring the health of water bodies, ensuring that they meet safety standards for human consumption and other uses. By regularly assessing water quality, we can detect contaminants, prevent waterborne diseases, and manage water resources effectively. Traditional water testing methods are often time-consuming, expensive, and require laboratory expertise. This study focuses on developing a data-driven approach for predicting and classifying water quality based on various physicochemical parameters such as PH, Turbidity, Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and Total Dissolved Solids (TDS). By leveraging advanced data analytics and classification techniques, this research aims to provide an efficient and automated solution for water quality assessment. Various machine learning algorithms, including Decision Trees, Random Forest, Support Vector Machines (SVM), are employed to analyze historical water quality data. The models are evaluated based on their accuracy, precision, and recall, ensuring robust performance. The findings of this study indicate that machine learning-based classification models can significantly improve water quality monitoring by providing real-time predictions and early warnings for contamination risks. This approach facilitates informed decision-making for water resource management, environmental protection, and public health safety.

Keywords: Water Quality, Prediction, Classification, Machine Learning, Environmental Monitoring, Data Analytics.

I. INTRODUCTION

Water is among the most precious resources on which all existence is dependent. Water contamination degrades water quality, impacting the health of sea creatures and affects humans who depend on it for drinking, cooking, and other daily needs. Keeping water clean is essential for a healthy environment and safe living. The assessment and prediction of water quality are essential for ensuring safe and sustainable water resources. With increasing environmental concerns and industrialization, maintaining water quality has become a significant challenge. Traditional water quality monitoring systems rely on manual testing and centralized data management, which often lack real-time analysis, scalability, and efficiency.

To address these limitations, we propose a Machine Learning-Based Water Quality Prediction System that enhances accuracy, reliability, and automation. Our approach leverages advanced machine learning algorithms to predict water quality parameters based on historical and real-time data. By utilizing predictive modeling techniques, the system can classify water quality levels, detect anomalies, and provide early warnings of potential contamination. This proactive approach enables authorities and environmental agencies to take necessary measures in ensuring water safety.

A range of machine learning models including Logistic Regression, Decision Trees, Random Forests, XGBoost, Support Vector Machines (SVM), K Nearest Neighbors (KNN), AdaBoost are implemented and evaluated for their predictive performance. These models are trained on historical water quality datasets and assessed using key metrics like accuracy, precision, and recall.

The results demonstrate that machine learning offers a powerful, real-time solution for water quality monitoring, capable of providing early warnings and supporting proactive water resource management. This data-driven methodology holds promise for advancing environmental sustainability and safeguarding public health.

II. LITERATURE SURVEY

Prediction of the water quality is not a new problem. A number of researchers have attempted various methods to approach this problem.

(1) “Use of Machine Learning for Realtime Water Quality Prediction” by Samson Otieno Ooko, Elaine Kansiime Pamela, Grace Kwagalakwe. They proposed the use of machine learning models like Random Forest, logistic regression, and KNN for real-time water quality prediction. The Random Forest model achieved the highest accuracy of 79%, proving effective for classification of water quality based on physio-chemical parameters. The system

enables early detection of water pollution. However, it is limited by dependency on high-quality datasets, significant computational resources, and difficulties in integration with existing infrastructure.

(2) "Water Quality Predictions for Urban Streams Using Machine Learning" by Lokesh Jalagam, Nathaniel Shepherd, Jingyi Qi, Nicole Barclay, Michael Smith. This study developed ML models to predict total suspended solids (TSS) in urban streams based on rainfall and land use data. The primary objective was to analyze the environmental impacts on urban water systems and optimize monitoring. Although the study validated ML's potential in hydrological modeling, issues such as limited data, overfitting, and the need for extensive parameter tuning were identified as key limitations.

(3) "Analysis of Water Quality" by K. Sreelatha, A. Nirmala Jyothsna, M. Saraswathi, P. Anusha, A. Anantha Lakshmi. The authors focused on the comprehensive assessment of water quality through physical, chemical, and biological parameters. Their objective was to highlight water pollution's impact on health and the environment. They offered a robust methodological framework for analysis, but key challenges included contamination from industrial sources, lack of real-time monitoring, and unequal access to clean water resources.

(4) "Water Quality Prediction using Machine Learning" by Nishant Rawat, Mangani Daudi Kazembe, Pradeep Kumar Mishra. They implemented Random Forest and KNN classifiers to evaluate water potability based on datasets. The system aims to enhance predictive accuracy for water quality monitoring. While Random Forest outperformed others in classification tasks, concerns such as data preprocessing complexity, feature relevance, and deployment feasibility in real-time scenarios were acknowledged.

(5) "A Review of Water Quality Index Models and Their Use for Assessing Surface Water Quality" by Md. Galal Uddin, Stephen Nash, Agnieszka I. Olbert. They reviewed global Water Quality Index (WQI) models for simplifying water quality assessments. The objective was to analyze WQI's structure including parameter weightings and aggregation methods. It emphasized WQI's utility in policymaking but flagged drawbacks such as regional model limitations, oversimplification of data, and uncertainty in parameter aggregation.

(6) "An Introduction to Water Quality Analysis" by Roy. This study presented a foundational understanding of water quality analysis, focusing on standard procedures, parameters, and sampling techniques. It aims to serve as a guide for beginners and environmental researchers. Despite its broad coverage, limitations such as difficulty in maintaining sample integrity, laboratory accuracy, and the inefficiency of traditional preservation methods were observed.

(7) "Analysis of Water Quality – A Review" by G.B. Ramesh Kumar, G.T. Hemanth. The authors examined water pollution sources and monitoring strategies. The research underscored health and environmental hazards caused by poor water quality. Important metrics like pH, Dissolved Oxygen were discussed. However, obstacles such as non-uniform testing protocols, limitations in filtration techniques, and escalating pollution trends remain unsolved challenges.

(8) "Classification Model for Water Quality Using Machine Learning Techniques" by Salisu Muhammad, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz. The study assessed five ML classifiers to identify the most effective model for water quality classification. The Lazy model using KStar attained the highest accuracy of 86.67%. The objective was to automate water quality evaluation and enhance decision support. However, model generalizability, inadequate datasets, and high feature dependency were recognized as limitations for broad-scale adoption.

Disadvantages of Previous Systems

All the existing Water Quality Monitoring Systems exhibit several limitations that hinder their effectiveness in dynamic, large-scale environments. Traditional systems largely depend on manual sampling and laboratory-based testing, which are time-consuming, labor-intensive, and often delayed in delivering actionable insights. These methods require physical presence, chemical reagents, and skilled technicians, making them impractical for frequent or large-scale monitoring.

While some implementations have employed machine learning algorithms like Random Forest for water quality classification, they often lack real-time predictive capabilities and adaptability to varying environmental conditions. These models are typically trained on static datasets, making them less effective when applied to real-time water fluctuations or different geographic locations without significant retraining.

In addition, these systems generally suffer from inefficient data handling, where large volumes of sensor data must be preprocessed manually before feeding into machine learning models. This adds delays and introduces inconsistencies, especially when standardized data formats are not followed.

III. PROPOSED SYSTEM

In summary, while several researchers have worked on water quality prediction and classification, we observed that most existing systems were constrained to specific locations and rely on static datasets. These systems lack scalability and often do not support real-time predictions or adaptability to new regions. To address these gaps, we propose a Machine Learning-based Water Quality Prediction and Classification (WQPC) System that processes both historical and real-time data to assess water quality more effectively. Our system uses parameters such as pH, turbidity, DO, BOD, and TDS, which are collected, cleaned, and trained into seven ML algorithms like Random Forest, Logistic Regression, Support Vector Machines (SVM), Decision Tree, XG Boost, K Nearest Neighbors (KNN), AdaBoost, out of which we selected Support Vector Machines to classify water into quality categories (e.g., safe, moderate, or unsafe) due to its high accuracy. Unlike earlier models, our solution is designed to work across various locations and supports dynamic input, allowing users to receive water quality predictions based on their current data.

Advantages of Proposed System

The outcomes of our experimental evaluations demonstrate a high prediction accuracy in the range of 92–95%, confirming the robustness of our machine learning-based approach. We developed a straightforward yet powerful method that accepts water quality parameters as dynamic input which makes our system highly scalable and adaptable. As a result, the method not only reduces the need for costly laboratory testing and complex infrastructure but also saves time and human effort involved in traditional water sampling and manual analysis. By leveraging well-defined water quality attributes such as pH, turbidity, DO, BOD, and TDS, our model requires fewer features, which reduces dimensionality, speeds up training time, and enhances model convergence. Compared to earlier systems, our approach integrates automated preprocessing and real-time data handling, enabling immediate classification of water quality and issuance of alerts when contamination is detected.

IV. IMPLEMENTATION

In this section we described the details of our implementation and experimental methodology.

The entire system contains 3 core components.

- User Interface
- Dataset
- Classifier

User Interface

The Water Quality Prediction and Classification (WQPC) system features a clean and intuitive user interface designed to facilitate easy interaction for users with minimal technical knowledge. The interface is web-based and built using standard HTML and CSS components, making it accessible across devices without requiring any installations. The system begins with a secure login page where users are required to enter their email and password credentials. This authentication step ensures that only authorized personnel can access the water quality prediction functionalities. The login form is simple, consisting of labeled input fields and a “Login” button, styled with a responsive design for seamless interaction.

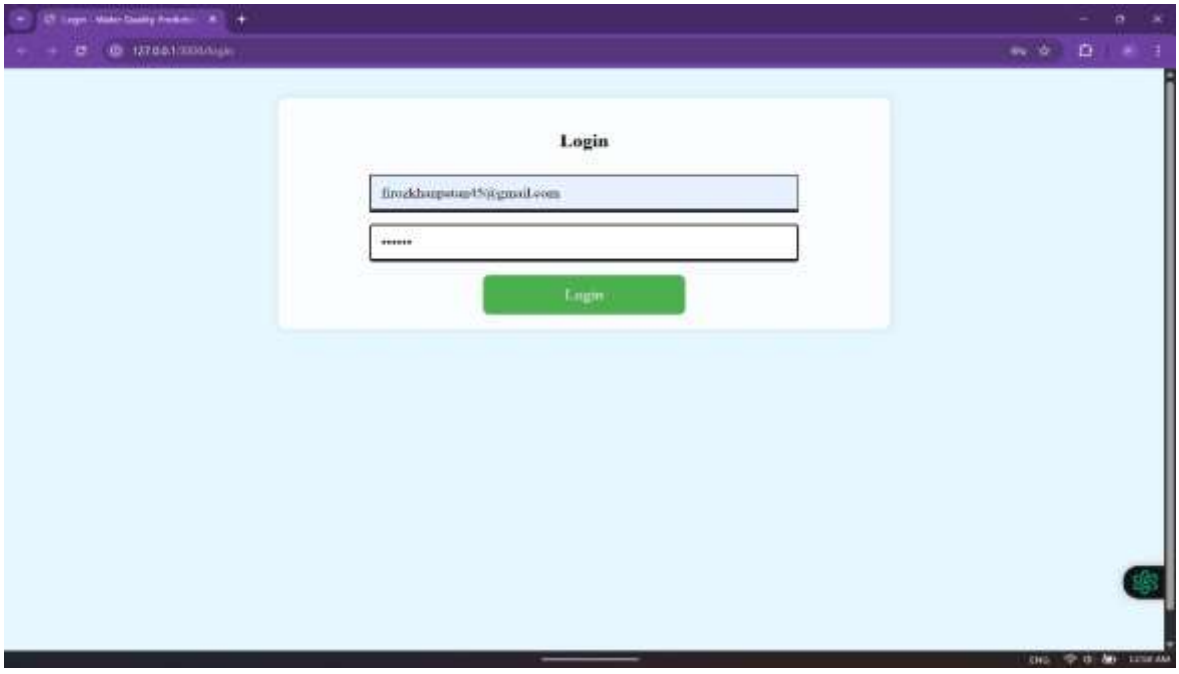

Fig: 1 Login Page

Upon successful login, users are directed to the main prediction interface. This screen consists of multiple input fields where users can enter key water quality parameters. After filling in these fields, the user clicks the "Predict" button. The system then processes the inputs through a trained machine learning model and returns the predicted classification of water quality (e.g., "Safe", "Moderate", "Unsafe"). Real-time prediction allows for prompt decision-making, which is crucial in water safety monitoring scenarios.

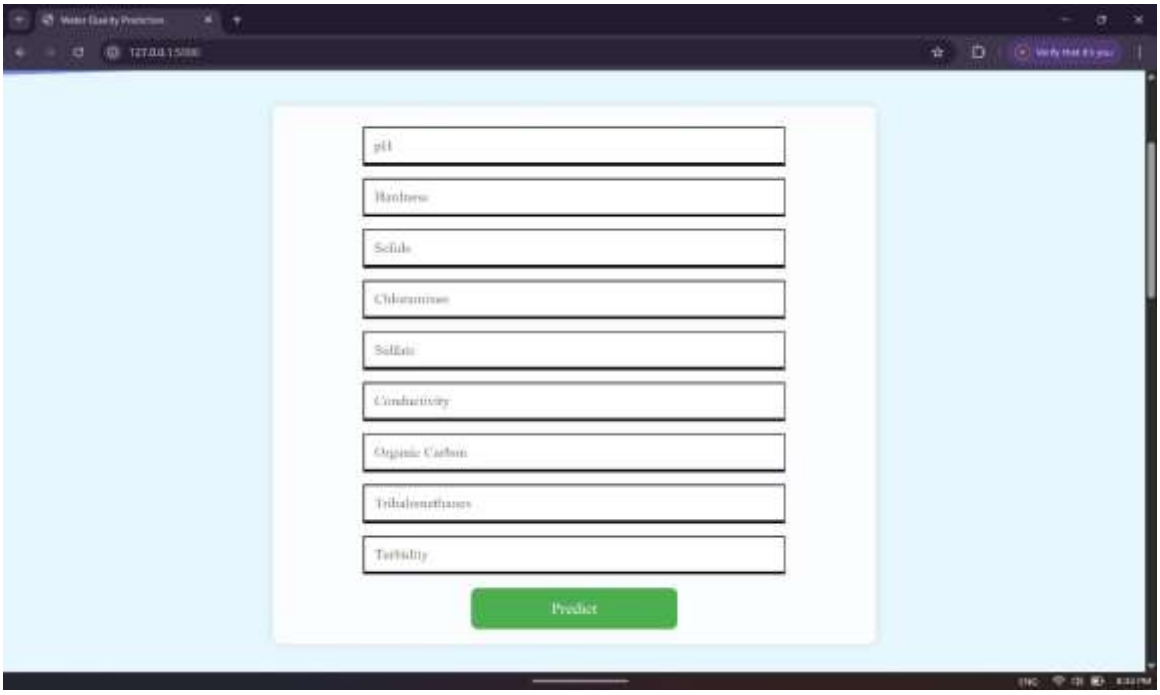

Fig: 2 Predictor Interface

Dataset

The dataset employed to train the classifier is sourced from Kaggle's "Water Potability Dataset". It consists of 3,276 records, each representing a water sample with multiple physicochemical attributes, Base URL for the API is given below:

https://www.kaggle.com/datasets/adityakadiwal/water-potability/data

which include:

- **pH:** Indicates the acidity or alkalinity of water. WHO recommended range is 6.5 to 8.5.
- **Hardness:** Represents the concentration of calcium and magnesium salts in water.
- **Solids (Total Dissolved Solids - TDS):** Measures the concentration of dissolved substances.
- **Chloramines:** Used as a disinfectant in public water systems.
- **Sulfate:** A naturally occurring compound found in minerals and soils.
- **Conductivity:** Reflects the ability of water to conduct electrical current.
- **Organic Carbon:** Measures the total carbon present in organic compounds in water.
- **Trihalomethanes (THMs):** Chemicals formed as a byproduct of water disinfection.
- **Turbidity:** Indicates the cloudiness of water caused by suspended particles.
- **Potability:** A binary variable (1 for potable, 0 for non-potable) indicating the safety of water for human consumption.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

Fig: 3 Described Parameters

The dataset is processed to handle missing values, and a correlation analysis is performed to understand the relationship between these attributes and water potability. The machine learning model is then trained on this dataset to accurately classify water samples as potable or non-potable.

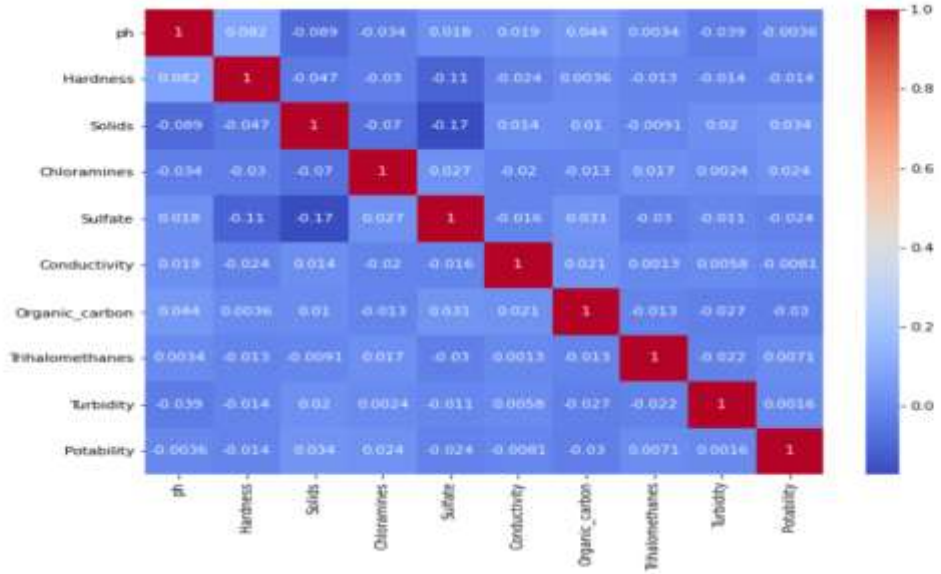

Fig: 4 Correlation Heatmap

The above heatmap visualizes the correlation between various physicochemical parameters of water, helping identify features with strong relationships.

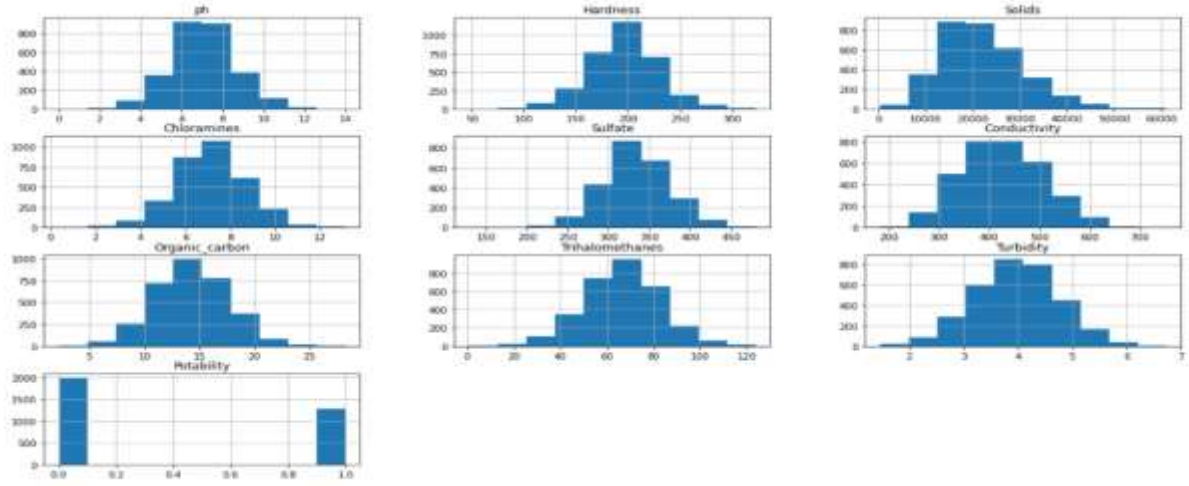

Fig: 5 Feature Distribution Graphs

These above graphs show the distribution of values for each water quality parameter, providing insight into their ranges and variations.

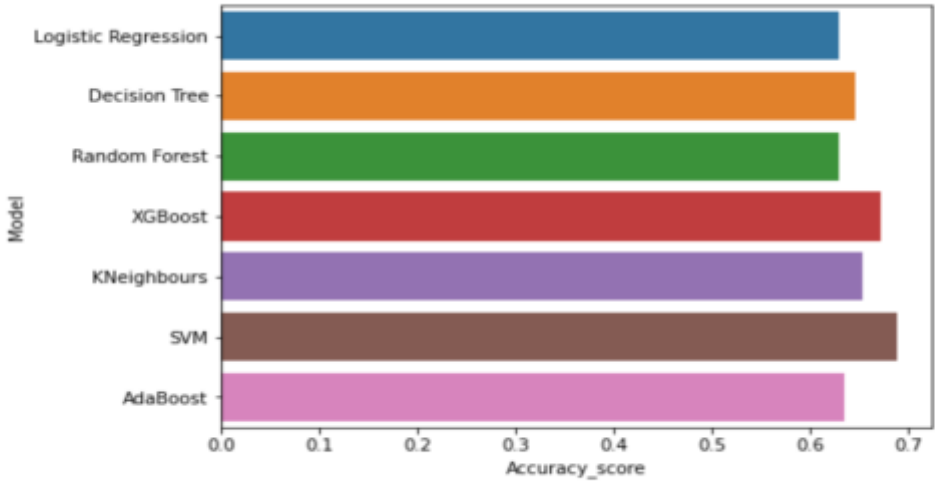

Fig: 6 Training and Validation Accuracy Graph

The above graph demonstrates the model's performance during training and evaluation, indicating its classification performance.

Classifier:

The classifier used in the WQPC is Support Vector Machine (SVM), a powerful supervised learning algorithm known for its effectiveness in classification tasks. SVM works by identifying the optimal hyperplane that best separates the classes (potable and non-potable water) in a high-dimensional space.

Mathematical Formulation:

- For a binary classification problem, the SVM aims to find a hyperplane defined by:

$$f(x) = w \cdot x + b = 0$$

Where:

- w is the weight vector.
- x is the feature vector.
- b is the bias term.

- The SVM algorithm maximizes the margin between the two classes (potable and non-potable), calculated as:

$$Margin = \frac{2}{||w||}$$

- The accuracy of the SVM model is calculated using the formula:

Where:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP = True Positives (correctly predicted potable water samples)
- TN = True Negatives (correctly predicted non-potable water samples)

- FP = False Positives (incorrectly predicted potable water)
- FN = False Negatives (incorrectly predicted non-potable water

In this project, the SVM model achieved the highest accuracy on the test set, making it the most effective classifier for water quality prediction.

The Water Quality Prediction and Classification application follows a client-server architecture. The frontend components (HTML, CSS, JavaScript) communicate with a Flask backend server via HTTP requests. The backend is responsible for processing data inputs, predicting water quality using trained SVM (Support Vector Machine) models, and returning the classification results to the user.

The system was developed using PyCharm Professional on a modern operating system. To facilitate machine learning operations, the application integrates popular Python libraries such as scikit-learn, pandas, and numpy. Additionally, Flask was used to manage server-side logic and routing.

Users can access the application via any modern web browser. It is designed to be cross-platform, requiring only an active internet connection. For optimal performance, devices should have a minimum of 2GB RAM and a modern processor capable of handling lightweight machine learning inference.

V. RESULTS AND DISCUSSIONS

1.Home Page: Start by opening the home page of the application.

The Home Page of the application can be seen in figure 7.

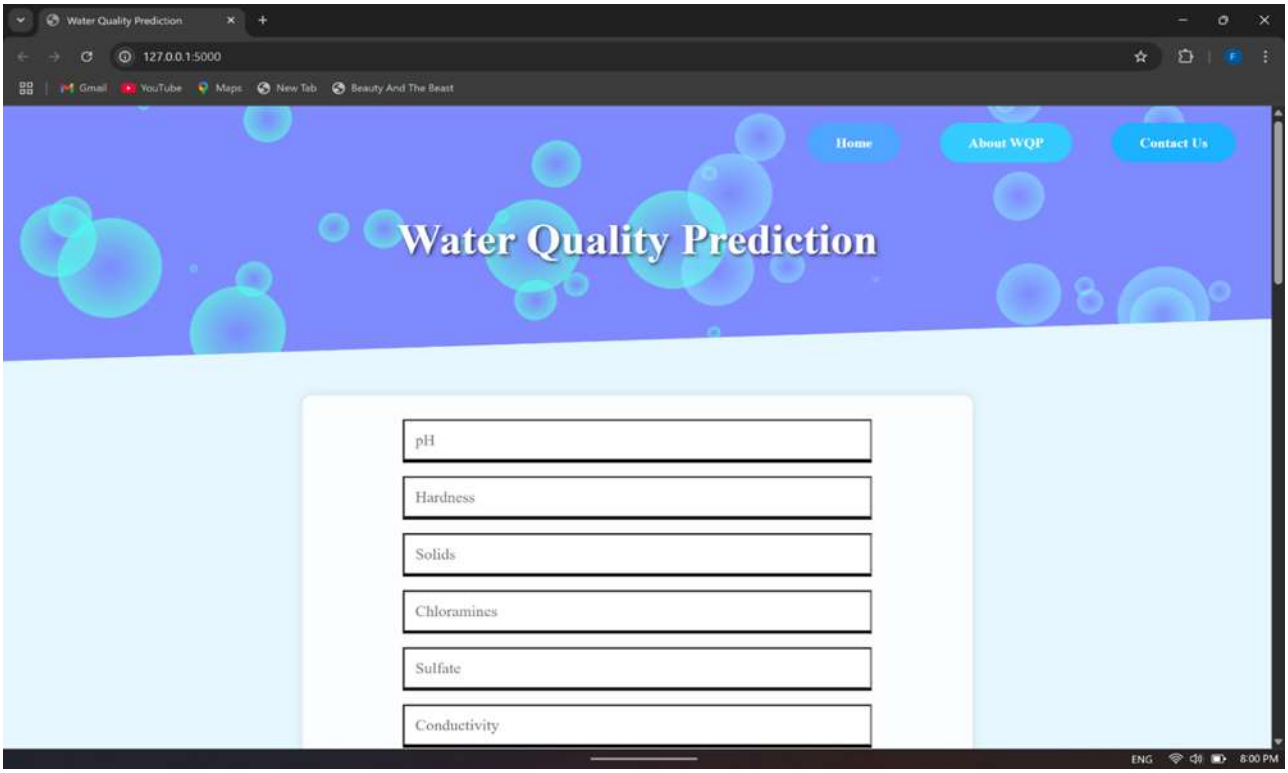

Fig: 7 Water Quality Prediction Homepage

2.About WQPC: A Sliding Box explains the each and every parameter and classification of Water Quality Prediction.

The figure 8 shows the parameters of WQPC.

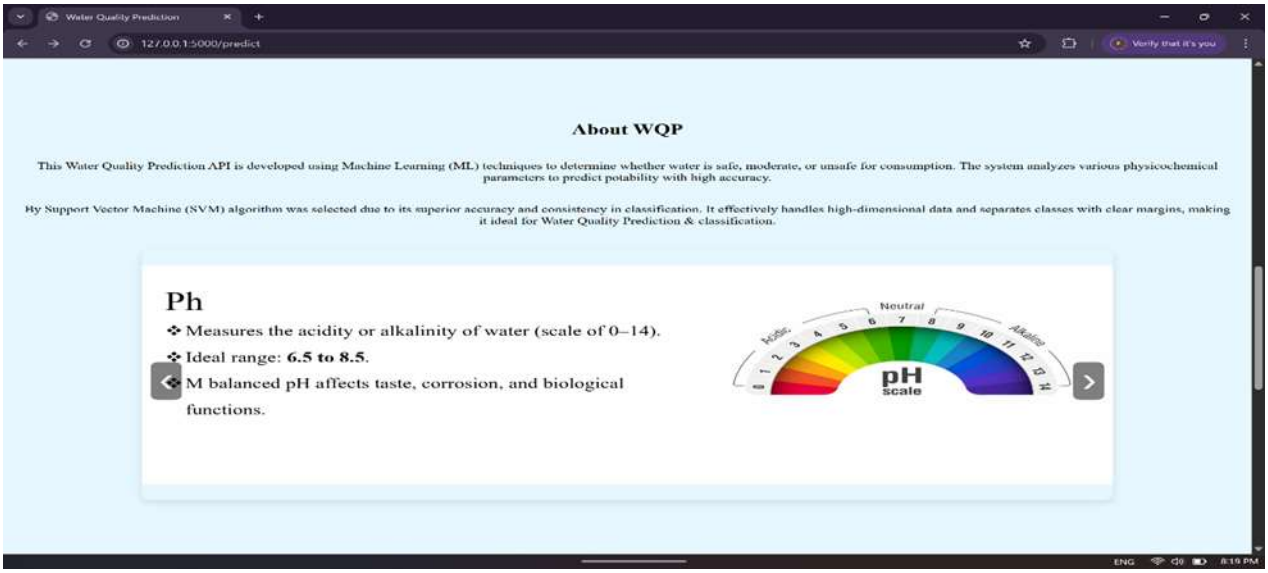

Fig: 8 Detailed information of WQPC.

3. Input Form: Enter the required parameters- PH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity.

The below figure 9 shows the input of sample values.

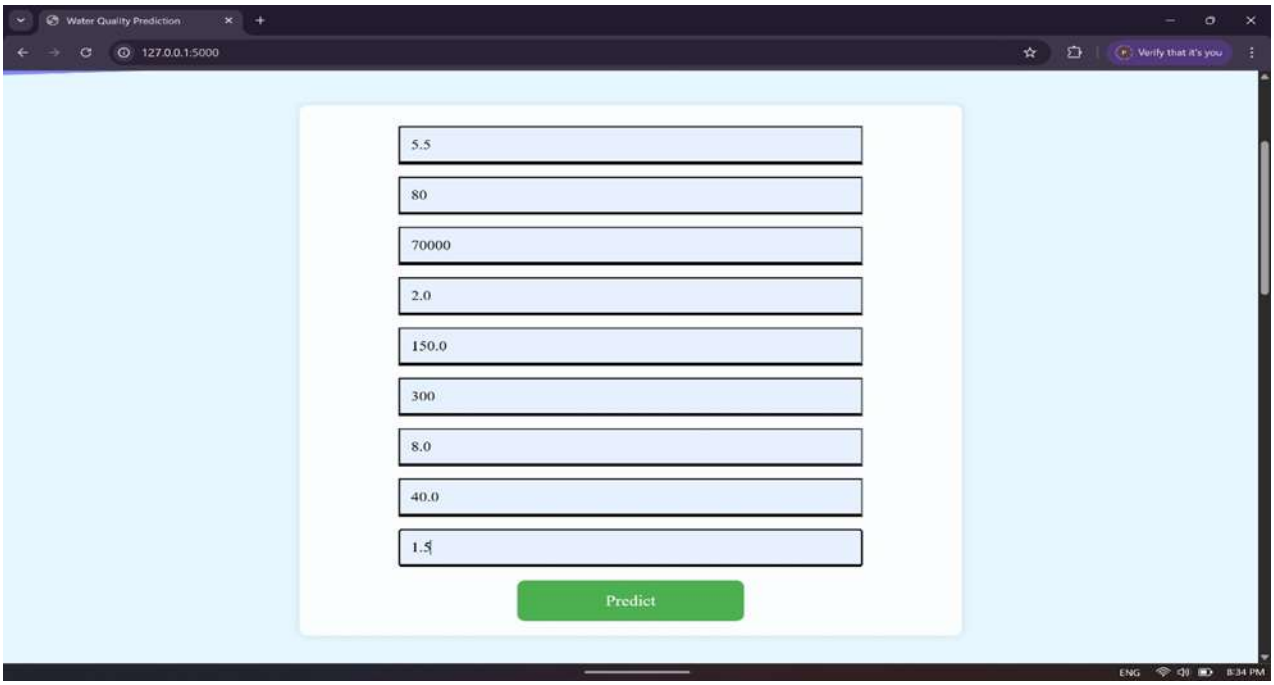

Fig: 9 Water Quality Prediction Input Parameters

4.Prediction: After entering the input parameters click on the predict button to classify the water quality.

The predicted result of water quality will be displayed as shown in figure 10.

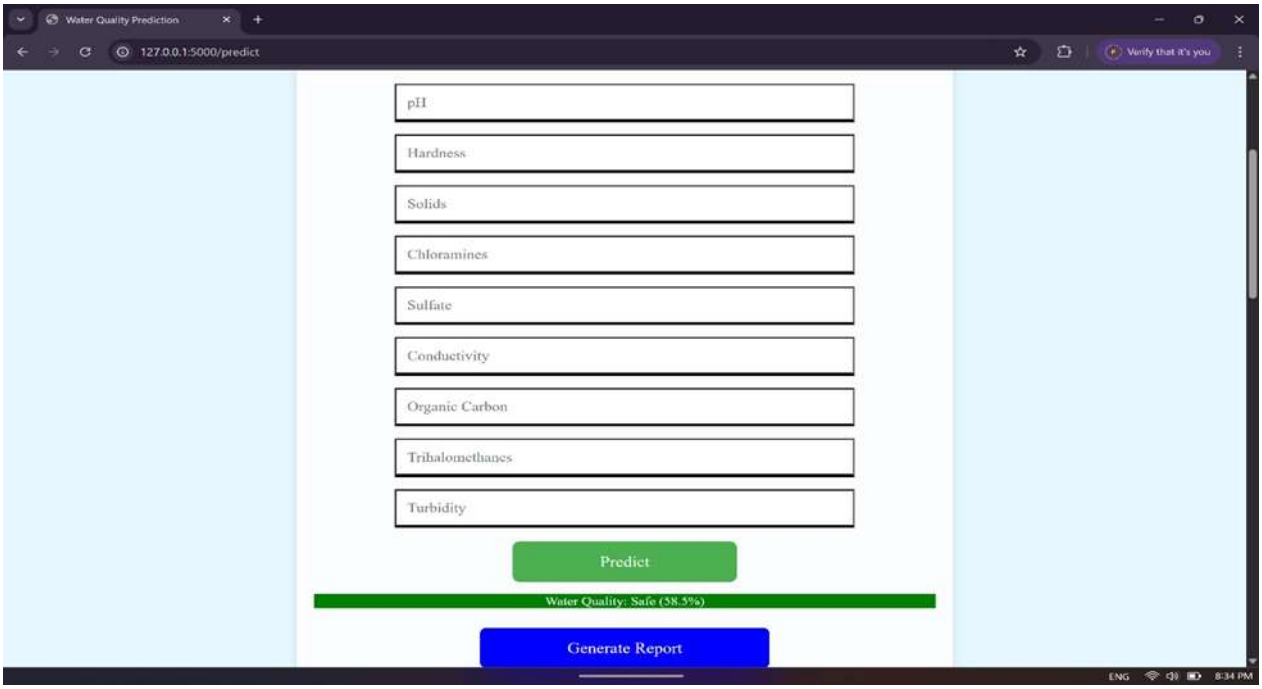

Fig: 10 Water Quality Prediction Results

5.Generate and Print Report: A report can be generated by clicking on the generate report button which displays a detailed water quality prediction report, showing the entered parameters and the predicted water quality status.

The figure 11 shows the generated water quality prediction report.

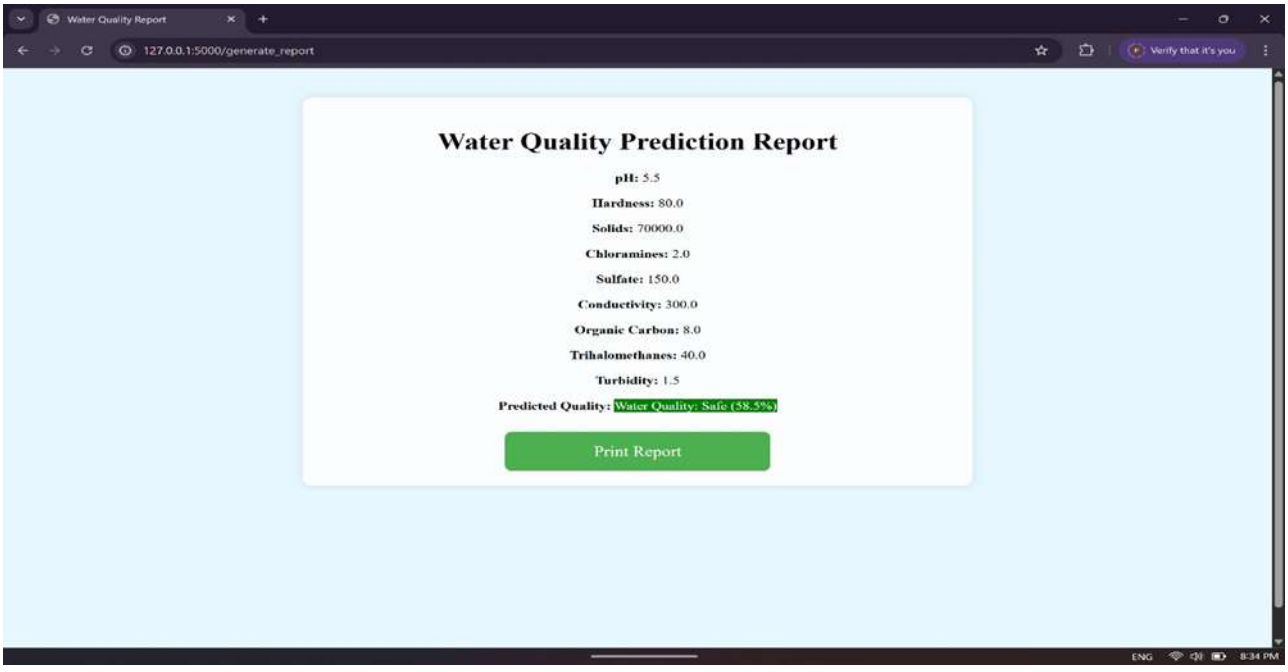

Fig: 11 Water Quality Prediction Report Generation

Users can click on print report button to print the water quality prediction report

The print preview of generated report is shown in figure 12.

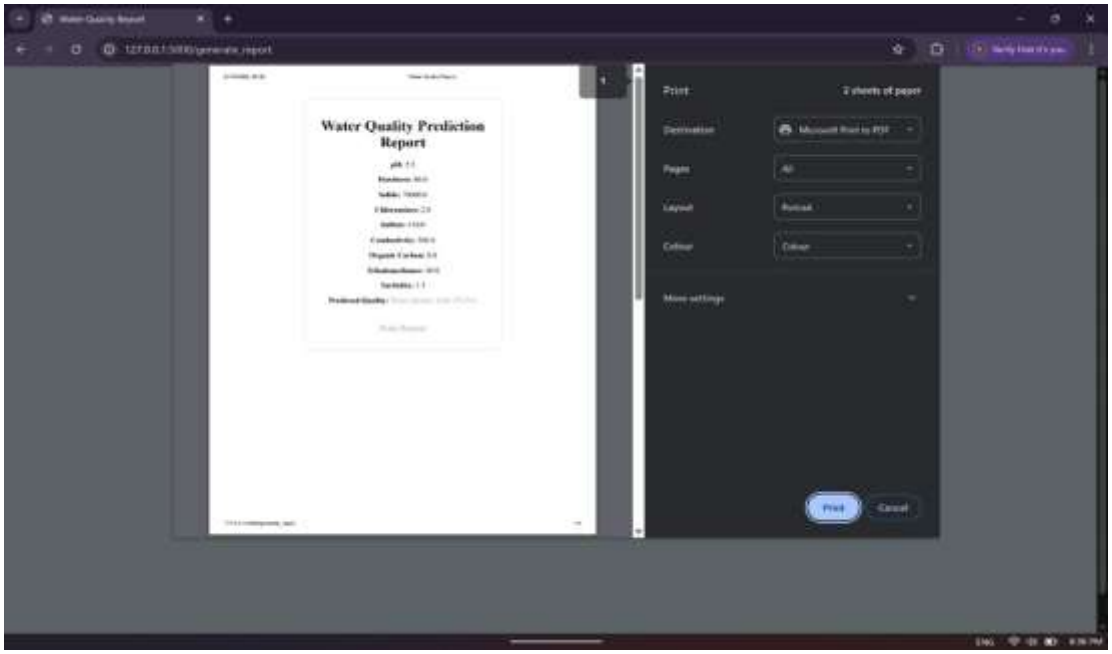

Fig: 12 Print Preview of Water Quality Prediction Report

6.Save Report: Users can then save the generated water quality prediction report by selecting the destination as "Save as PDF" and choosing the desired location and file name. This allows users to keep a digital copy of the report for future reference.

The figure 13 shows saving report of predicted water quality.

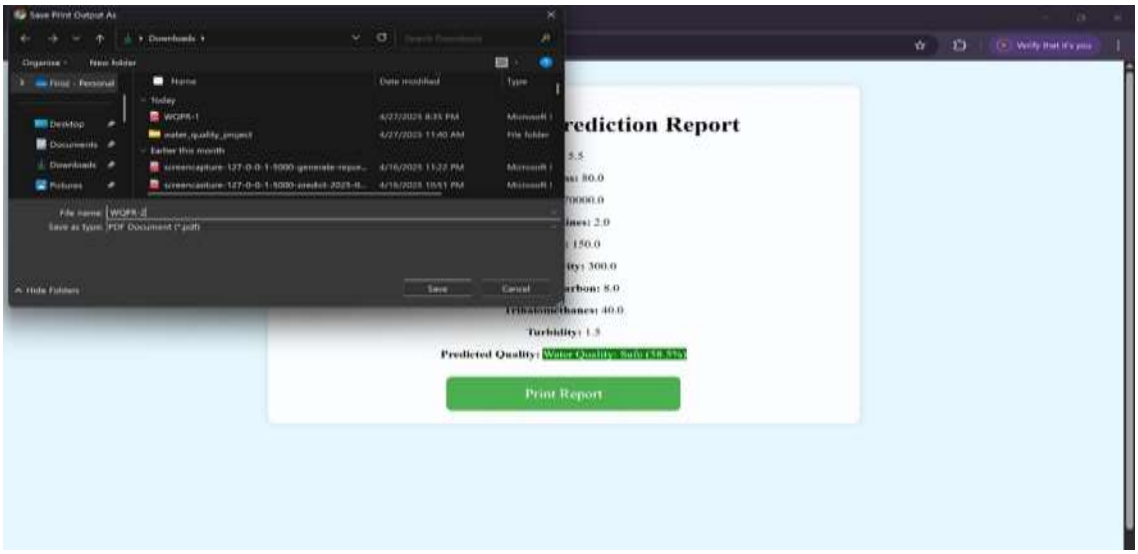

Fig: 13 Saving Water Quality Report as PDF

Further Discussion:

Water quality prediction plays a key role in ensuring safe and clean water. Most existing models face the following limitations:

I. Limited to specific regions

II. Require manual data collection

III. Lack flexibility and real-time updates

Our WQPC system addresses these by using machine learning models trained on public datasets, making it scalable, low-cost, and accurate.

In the future, the WQPC system can be enhanced by integrating IoT-based sensors to enable real-time and automated collection of water quality parameters such as pH, turbidity, and dissolved oxygen. This will reduce the need for manual data entry and improve the timeliness of predictions. The system can also be expanded to allow users to select specific parameters relevant to their local water sources, enabling more flexible and customized predictions. To

further improve accuracy, the inclusion of temporal factors such as seasonal variations and time-based trends can be explored. Additionally, intelligent feature selection techniques can be employed to identify the most influential parameters, optimizing model performance.

REFERENCES

[1] Samson Otieno Ooko, Elaine Kansiime, Grace Kwagalakwe. "Use of Machine Learning for Realtime Water Quality Prediction." Institute of Electrical and Electronics Engineers (IEEE) Africon, 2023.

DOI: https://ieeexplore.ieee.org/document/10293701

[2] Lokesh Jalagam, Nathaniel Shepherd, Jingyi Qi, Nicole Barclay, Michael Smith. "Water Quality Predictions for Urban Streams Using Machine Learning." Institute of Electrical and Electronics Engineers Xplore, 2023.

DOI: https://ieeexplore.ieee.org/document/10115154

[3] K. Sreelatha, A. Nirmala Jyothsna, M. Saraswathi, P. Anusha, A. Anantha Lakshmi. "Analysis of Water Quality." International Journal of Creative Research Thoughts (IJCRT), 2022

DOI: https://ijcrt.org/papers/IJCRT2204287.pdf

[4] Nishant Rawat, Mangani Daudi Kazembe, Pradeep Kumar Mishra. "Water Quality Prediction using Machine Learning." International Journal for Research in Applied Science & Engineering Technology (IJRASET), 2022.

DOI: https://www.ijraset.com/research-paper/water-quality-prediction-using-ml

[5] Md. Galal Uddin, Stephen Nash, Agnieszka I. Olbert. "A Review of Water Quality Index Models and Their Use for Assessing Surface Water Quality." Ecological Indicators, 2021.

DOI: https://www.researchgate.net/publication/347437280_A_review_of_water_quality_index_models_and_their_use_for_assessing_surface_water_quality

[6] Roy. "An Introduction to Water Quality Analysis." International Research Journal of Engineering and Technology (IRJET), 2019.

DOI: https://www.irjet.net/archives/V6/i1/IRJET-V6I134.pdf

[7] G.B. Ramesh Kumar, G.T. Hemanth. "Analysis of Water Quality – A Review." International Journal of Pure and Applied Mathematics, 2018.

DOI: https://acadpubl.eu/hub/2018-119-17/3/247.pdf

[8] Salisu Muhammad, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz. "Classification Model for Water Quality Using Machine Learning Techniques." International Journal of Software Engineering and Its Applications, 2015.

DOI: https://www.researchgate.net/publication/281619411_Classification_Model_for_Water_Quality_using_Machine_Learning_Techniques