Intelligent Water Quality Prediction Using Machine Learning Models for Potability Assessment

KOTIKALAPUDI MOHAN SATYA BHAGAVAN

PG Scholar. Department of MCA, DNR College, Bhimavaram, Andhra Pradesh

A.Durga Devi

(Assistant Professor), Master of Computer Applications, DNR College, Bhimavaram, Andhra Pradesh

ABSTRACT

Water quality assessment is a critical aspect of environmental monitoring and public health management. With increasing industrialization, urbanization, and agricultural activities, water bodies are continuously exposed to pollutants, making it essential to monitor and evaluate water quality efficiently. Traditional methods of water quality analysis rely on laboratory testing, which is time-consuming, expensive, and not suitable for real-time monitoring. To overcome these limitations, this research proposes an intelligent water quality prediction system using machine learning techniques for potability assessment.The proposed system leverages data-driven approaches to classify water samples as potable or non-potable based on physicochemical parameters. The implementation is developed using Python and integrates machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). A graphical user interface is designed using a GUI framework to facilitate user interaction, enabling easy dataset upload, preprocessing, model execution, and result visualization.The system begins with data acquisition, where water quality datasets containing parameters such as pH, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity are used. Preprocessing techniques are applied to handle missing values and normalize the dataset. The data is then divided into training and testing sets to evaluate model performance.Feature extraction and encoding techniques are used to transform the dataset into a suitable format for machine learning models. The SVM and Random Forest algorithms are employed for classification, providing robust performance for structured data. Additionally, an Artificial Neural Network is implemented with multiple hidden layers to capture complex nonlinear relationships among water quality parameters. The ANN model is trained over multiple epochs, and its performance is evaluated using accuracy metrics.The system also includes a visualization module that plots accuracy and loss graphs, providing insights into model performance during training. The prediction module allows users to input new data and obtain real-time classification results, indicating whether the water is safe for consumption.Experimental results demonstrate that the proposed system achieves high classification accuracy and effectively distinguishes between potable and non-potable water samples. Among the implemented models, the neural network shows improved performance in capturing complex relationships, while Random Forest provides stable and reliable results.

The proposed framework offers a scalable, efficient, and cost-effective solution for water quality prediction. It reduces dependency on manual testing and enables rapid decision-making. This system can be extended for real-time monitoring using IoT devices and integrated into smart environmental management systems. The research contributes to the development of intelligent water quality assessment tools, promoting sustainable resource management and public health protection.

Keywords: Water Quality Prediction, Machine Learning, Potability Classification, Artificial Neural Networks, Random Forest, Support Vector Machine, Environmental Monitorin

I. INTRODUCTION

Water is one of the most essential natural resources, playing a vital role in human survival, agriculture, industry, and ecosystem balance. Ensuring the availability of safe and clean drinking water is a major global challenge, especially in developing countries. Water contamination caused by industrial discharge, agricultural runoff, and urban waste has significantly degraded water quality, leading to serious health risks and environmental issues.

Traditional water quality assessment methods involve laboratory-based analysis, where water samples are collected and tested for various parameters. While these methods provide accurate results, they are time-consuming, expensive, and not suitable for continuous monitoring. In many cases, delays in analysis can lead to the consumption of contaminated water, posing health hazards. Therefore, there is a need for efficient and real-time water quality prediction systems.Advancements in machine learning have opened new possibilities for environmental monitoring and data analysis. Machine learning models can analyze large datasets, identify patterns, and make predictions with high accuracy. These models are particularly useful in scenarios where relationships between variables are complex and nonlinear, as in the case of water quality parameters.This research focuses on developing a machine learning-based system for predicting water quality and determining its potability. The system utilizes multiple algorithms, including Support Vector Machine, Random Forest, and Artificial Neural Networks, to classify water samples based on physicochemical attributes. Each algorithm offers unique advantages, with SVM providing strong classification capabilities, Random Forest ensuring robustness, and ANN capturing complex patterns.The motivation behind this work is to create an intelligent, efficient, and scalable solution for water quality assessment. By automating the classification process, the system reduces reliance on manual testing and enables faster decision-making. Additionally, the integration of a graphical user interface enhances usability, allowing users to interact with the system easily.The key contributions of this research include the development of a multi-model classification framework, implementation of a user-friendly interface, and evaluation of model performance using real datasets. The study demonstrates that machine learning techniques can significantly improve the efficiency and accuracy of water quality prediction systems.

II. LITERATURE SURVEY (WITH EXISTING METHODS)

Water quality prediction has been extensively studied using various computational and statistical approaches. Traditional methods primarily rely on laboratory analysis and statistical modeling techniques. These approaches involve regression models and empirical formulas to estimate water quality parameters. While effective for small datasets, they often fail to capture complex relationships in large-scale data.Machine learning techniques have gained popularity due to their ability to handle large datasets and model nonlinear relationships. Early studies utilized algorithms such as Decision Trees and Naïve Bayes for water quality classification. These methods provided moderate accuracy but were limited by their inability to generalize across diverse datasets.Support Vector Machines have been widely used for water quality prediction due to their effectiveness in high-dimensional spaces. SVM models create optimal decision boundaries and perform well in classification tasks. However, they require careful parameter tuning and may not scale efficiently for very large datasets.Random Forest, an ensemble learning technique, has shown promising results in environmental data analysis. By combining multiple decision trees, it improves accuracy and reduces overfitting. Studies have demonstrated that Random Forest performs well in predicting water quality parameters and classifying water samples.Artificial Neural Networks have been extensively applied in recent years for water quality prediction. ANN models are capable of capturing complex nonlinear relationships between input parameters and output classes. Deep learning approaches further enhance prediction accuracy by using multiple hidden layers. However, these models require large datasets and computational resources.Recent research has also explored hybrid models that combine multiple algorithms to improve performance. Additionally, big data techniques and IoT-based monitoring systems have been integrated with machine learning models for real-time water quality assessment.Despite these advancements, challenges remain in handling missing data, ensuring scalability, and achieving consistent accuracy across different datasets. This research builds upon existing methods by implementing multiple machine learning models within a unified framework, providing a comprehensive solution for water quality prediction.

III. EXISTING SYSTEM

Existing water quality assessment systems primarily rely on manual sampling and laboratory testing. These methods involve collecting water samples and analyzing them for various chemical and physical parameters. While accurate, these processes are time-consuming, costly, and not suitable for real-time monitoring.Statistical models have been used to predict water quality based on historical data. However, these models often assume linear relationships between variables, which limits their ability to capture complex interactions among parameters. As a result, their prediction accuracy is often insufficient for practical applications.Some systems have adopted machine learning techniques, but they typically use a single algorithm, which may not provide optimal performance across different datasets. Additionally, many existing systems lack user-friendly interfaces, making them difficult to use for non-technical users.

Another limitation is the inability to handle missing or noisy data effectively. Real-world datasets often contain incomplete or inconsistent information, which can negatively impact model performance. Furthermore, many systems are not scalable and cannot handle large datasets efficiently.These limitations highlight the need for an advanced system that combines multiple machine learning models, handles real-world data challenges, and provides accurate and efficient water quality predictions.

IV. PROPOSED METHOD

The proposed system introduces an intelligent framework for water quality prediction using multiple machine learning algorithms. Unlike traditional systems, this approach integrates Support Vector Machine, Random Forest, and Artificial Neural Networks to improve classification accuracy and reliability.The system begins with dataset upload through a graphical user interface, allowing users to easily input water quality data. Preprocessing techniques are applied to clean the dataset and remove missing values. The data is then shuffled and split into training and testing sets to ensure unbiased evaluation.

Each machine learning model is trained using the processed dataset. The SVM model provides strong classification capabilities, while the Random Forest model enhances robustness through ensemble learning. The Artificial Neural Network captures complex nonlinear relationships, improving overall prediction accuracy.The system also includes a visualization module that displays accuracy and loss graphs, helping users understand model performance. A prediction module allows users to input new data and receive instant classification results.The proposed system addresses the limitations of existing approaches by providing a scalable, efficient, and user-friendly solution. It improves accuracy, reduces manual effort, and enables real-time water quality prediction, making it suitable for modern environmental monitoring applications.

V. IMPLEMENTATION

The implementation of the proposed water quality prediction system is carried out using Python, integrating machine learning libraries and a graphical user interface for user interaction. The system is designed to provide an end-to-end solution for water potability classification, starting from dataset upload to final prediction and visualization.The implementation begins with the development of a user interface using a GUI framework. This interface allows users to upload water quality datasets, preprocess data, run machine learning models, and visualize results. The dataset typically consists of physicochemical parameters such as pH, hardness, dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, which are essential indicators of water quality.Once the dataset is uploaded, preprocessing is performed to clean and structure the data. Missing values are removed, and the dataset is shuffled to eliminate bias. The features (input variables) and labels (output classes) are separated, and categorical encoding is applied to convert class labels into machine-readable format. The dataset is then divided into training and testing sets using a standard split ratio to evaluate model performance.The system implements multiple machine learning models, including Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Networks

(ANN). The SVM model is trained to find optimal decision boundaries for classification, while the Random Forest model uses ensemble learning to improve accuracy and reduce overfitting. The ANN model consists of multiple dense layers with activation functions, enabling it to capture complex nonlinear relationships among input features.During training, the ANN model is optimized using backpropagation and gradient descent techniques. The training process is monitored using accuracy and loss metrics, which are plotted graphically to visualize model performance. Research shows that neural networks and ensemble models significantly improve prediction accuracy compared to traditional approaches .The prediction module allows users to input new data and obtain classification results indicating whether water is potable or non-potable. The system processes the input data through the trained model and displays the results in the interface.Overall, the implementation demonstrates a scalable and efficient system capable of handling real-world datasets. The integration of multiple models ensures robustness, while the GUI enhances usability, making the system suitable for practical environmental monitoring applications.

VI. ALGORITHMS

The proposed system utilizes a combination of machine learning algorithms to classify water quality based on physicochemical parameters.

Step 1: Data Collection

Water quality data is collected from datasets containing multiple parameters affecting potability.

Step 2: Data Preprocessing

The dataset is cleaned by removing missing values and normalizing features. Data is shuffled to avoid bias during training.

Step 3: Feature Selection

Relevant features such as pH, turbidity, and dissolved solids are selected as input variables for the model.

Step 4: Data Splitting

The dataset is divided into training and testing sets to evaluate model performance.

Step 5: Model Training

- **SVM Algorithm:** Constructs a hyperplane to separate classes.
- **Random Forest:** Uses multiple decision trees for classification.
- **ANN:** Learns complex relationships using hidden layers.

Step 6: Prediction

The trained model processes new input data and predicts whether water is potable or non-potable.

Step 7: Evaluation

Model performance is evaluated using accuracy metrics. Studies show that ensemble and neural models can achieve high accuracy in water quality prediction tasks .

Step 8: Visualization

Accuracy and loss graphs are generated to analyze model performance.

This multi-model approach improves reliability and ensures accurate classification.

VII. SYSTEM DESIGN

The system design follows a modular architecture, ensuring scalability, flexibility, and efficient processing of water quality data.

1. Input Layer

The system accepts water quality datasets containing physicochemical parameters. Users can upload datasets through the graphical interface.

2. Preprocessing Module

This module cleans and prepares data by removing missing values, normalizing features, and encoding labels. Proper preprocessing improves model accuracy and efficiency.

3. Feature Extraction Module

Relevant features are extracted from the dataset and converted into numerical form suitable for machine learning algorithms.

4. Machine Learning Module

This is the core component of the system. It includes:

- Support Vector Machine for classification
- Random Forest for ensemble learning
- Artificial Neural Network for deep learning

Each model processes the dataset and generates predictions based on learned patterns.

5. Training and Testing Module

The dataset is divided into training and testing sets. Models are trained on the training data and evaluated on the testing data to measure performance.

6. Visualization Module

This module generates graphs showing accuracy and loss during training. Visualization helps in understanding model behavior and performance trends.

7. Prediction Module

Users can input new data to predict water quality. The system processes the data through trained models and displays results.

8. Backend Processing

The backend manages system operations, including data handling, model execution, and result generation. It ensures smooth communication between modules.

9. Output Layer

The output layer displays classification results, indicating whether water is potable or non-potable.

10. Scalability

The system is designed to handle large datasets and can be extended to include advanced models such as deep learning and automated machine learning. Recent studies highlight the importance of scalable ML systems for real-time water quality monitoring .

The overall system design ensures efficient data processing, accurate prediction, and user-friendly interaction.

SYSTEM DESIGN IMAGES

Water is an important and essential element for the life on earth. Due to the growth of population and industrialization the water resources become more polluted. Waste disposal from industry, human wastes, automobile wastes, agricultural runoff from farmlands containing chemical factors, unwanted nutrients, and other wastes from point and non-point source flow to water bodies, which affects the quality of the water resources. etc. The increase in pollution influences the quantity and quality of water, which results high risk on health and other issues for human as well as for living organisms on the planet. Hence, evaluating and monitoring the quality of water, and its prediction become crucial and applicable area for research in the current scenario

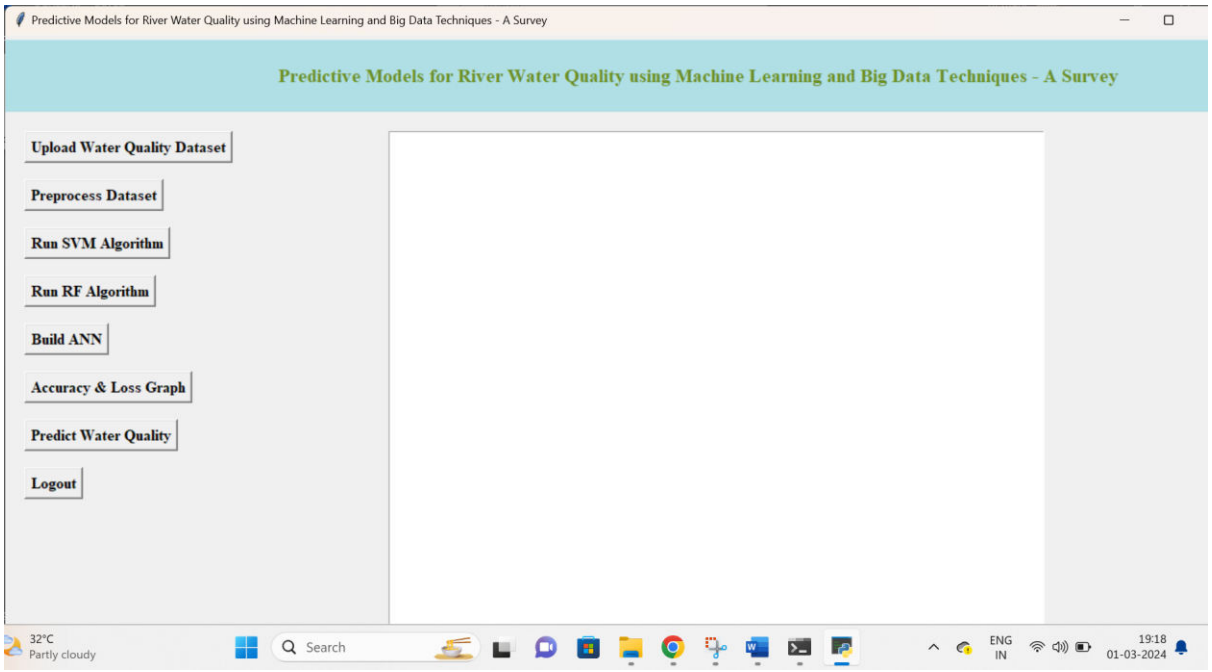
We used SVM ,RF and NN for prediction of water potability using different conditions.

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
	204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3.716089	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
	118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.97664	4.370562	0
7.360648	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0

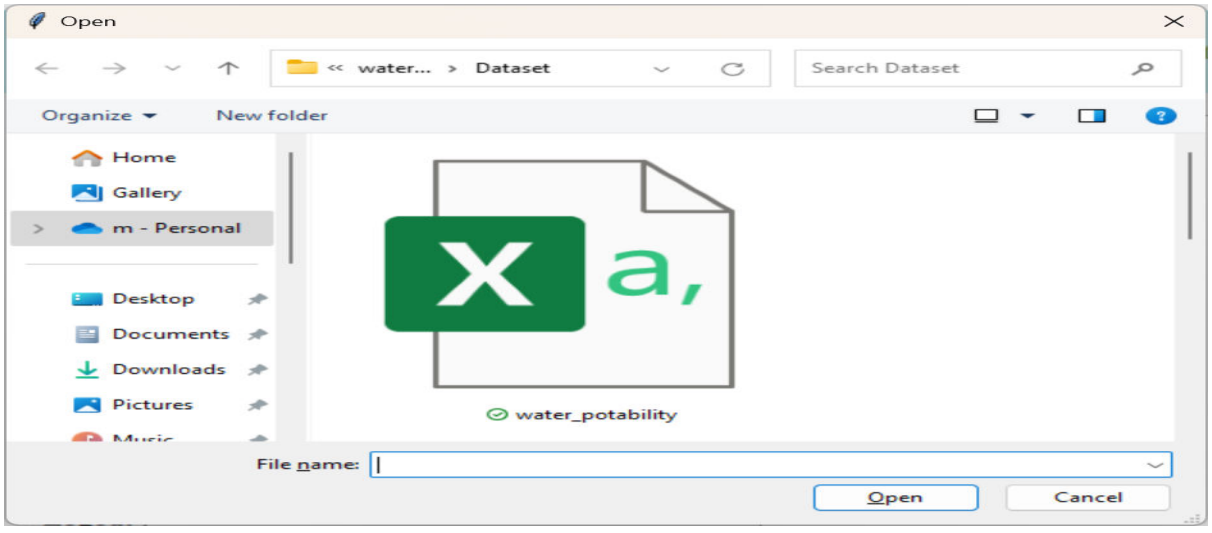
In above dataset all bold names are the dataset column names and all integer values are the dataset values.

SCREEN SHOTS

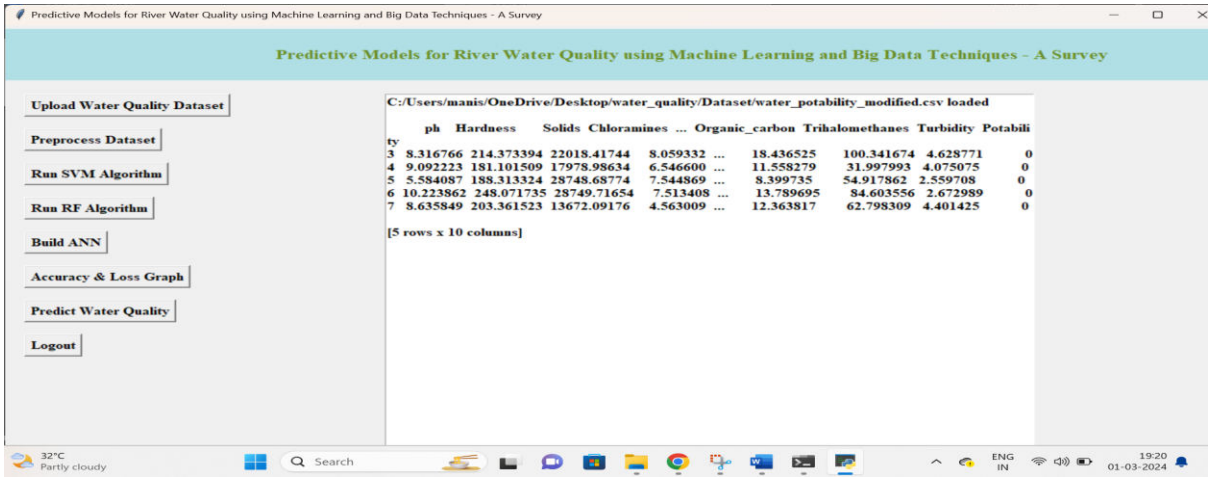
To run project double click on ‘run.bat’ file to get below screen



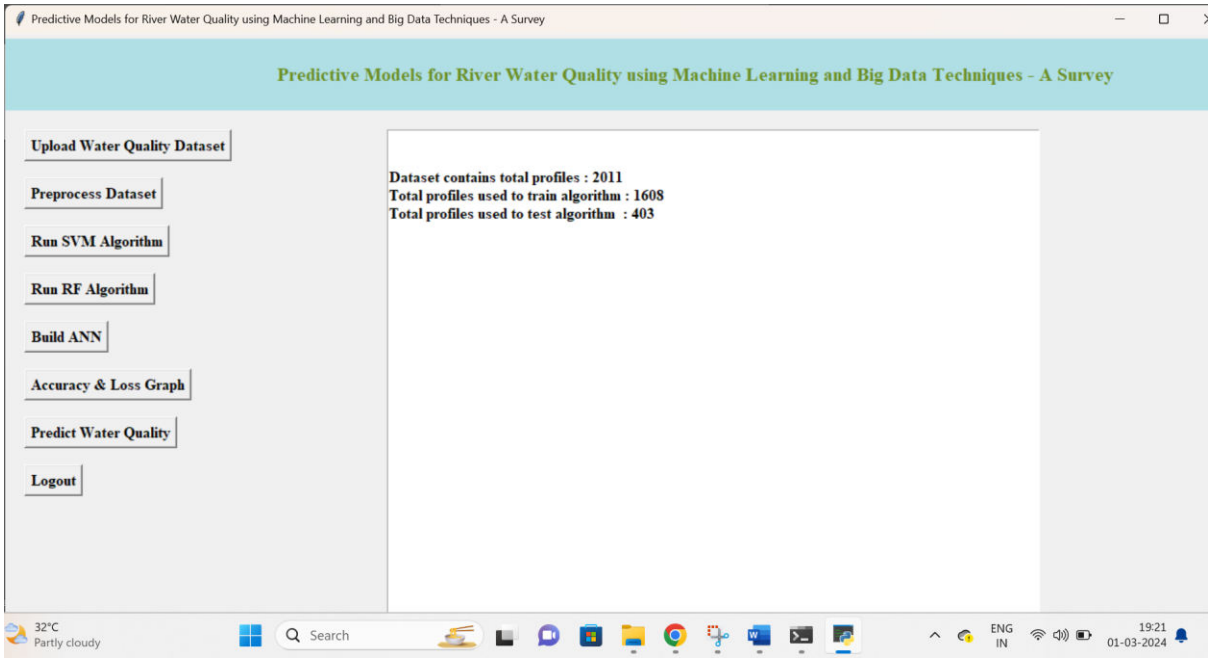
In above screen click on ‘Upload water quality Dataset’ button and upload dataset



In above screen selecting and uploading ‘dataset.csv’ file and then click on ‘Open’ button to load dataset and to get below screen

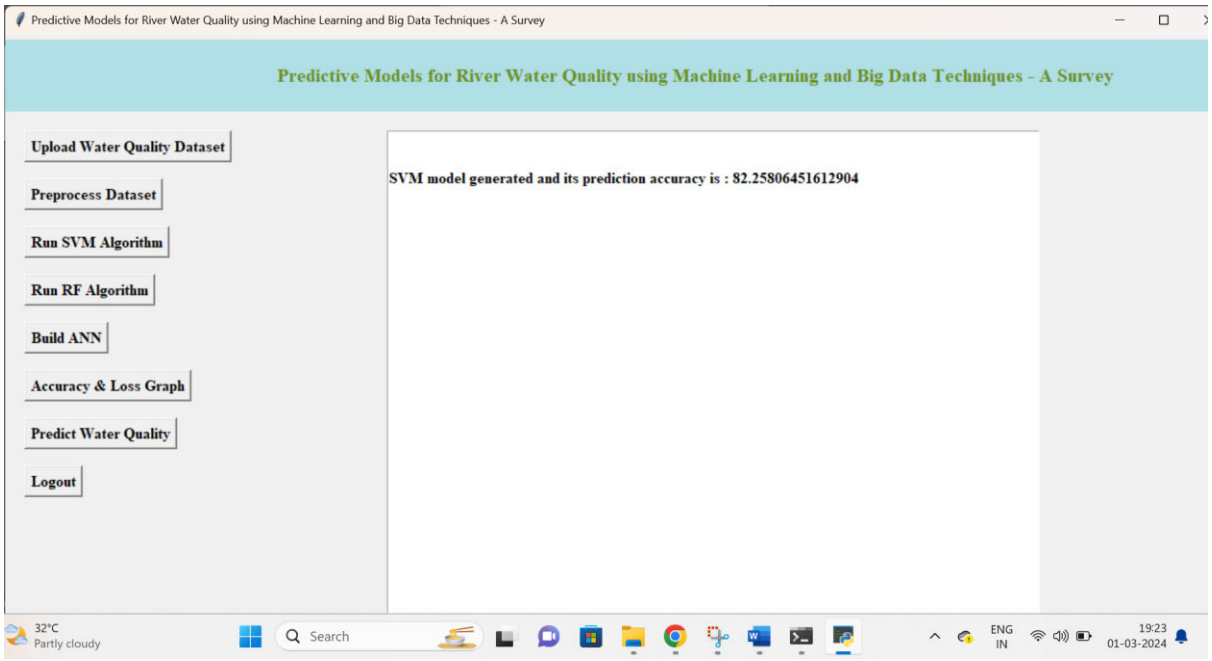


In above screen dataset loaded and displaying few records from dataset and now click on ‘Preprocess Dataset’ button to remove missing values and to split dataset into train and test part



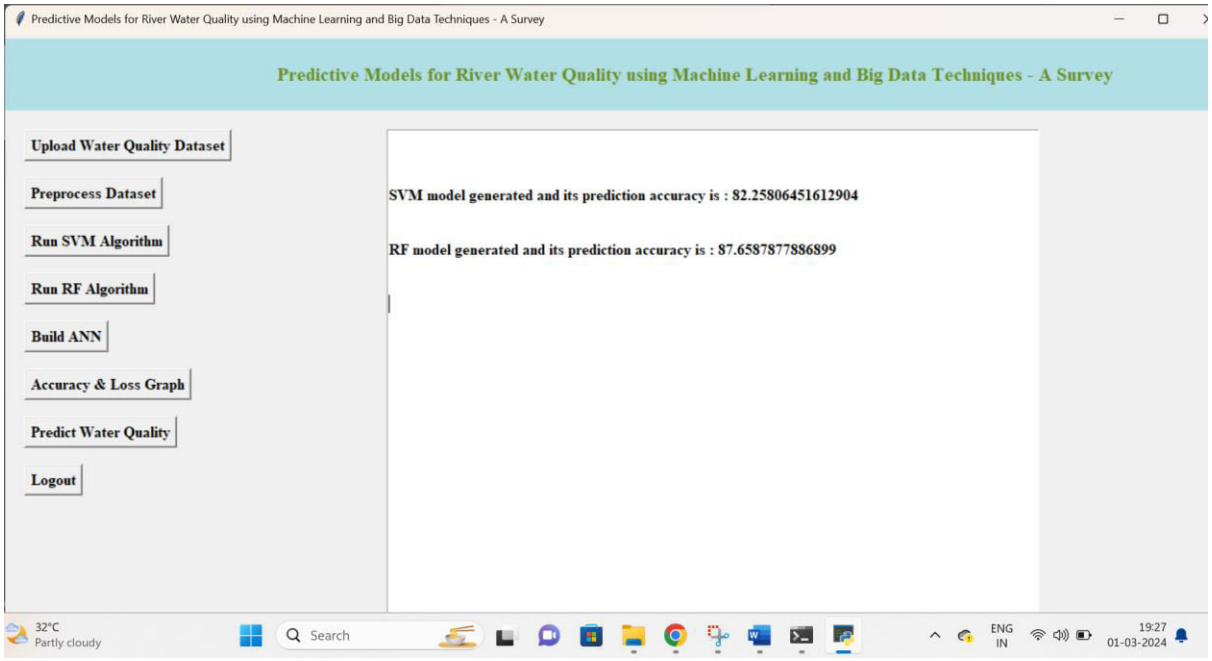
In above screen we can see dataset contains total 2011 records and application using 1608 records for training and 403 records to test ML algorithms and now dataset is ready and now click on ‘Run SVM Algorithm’ button to SVM algorithm

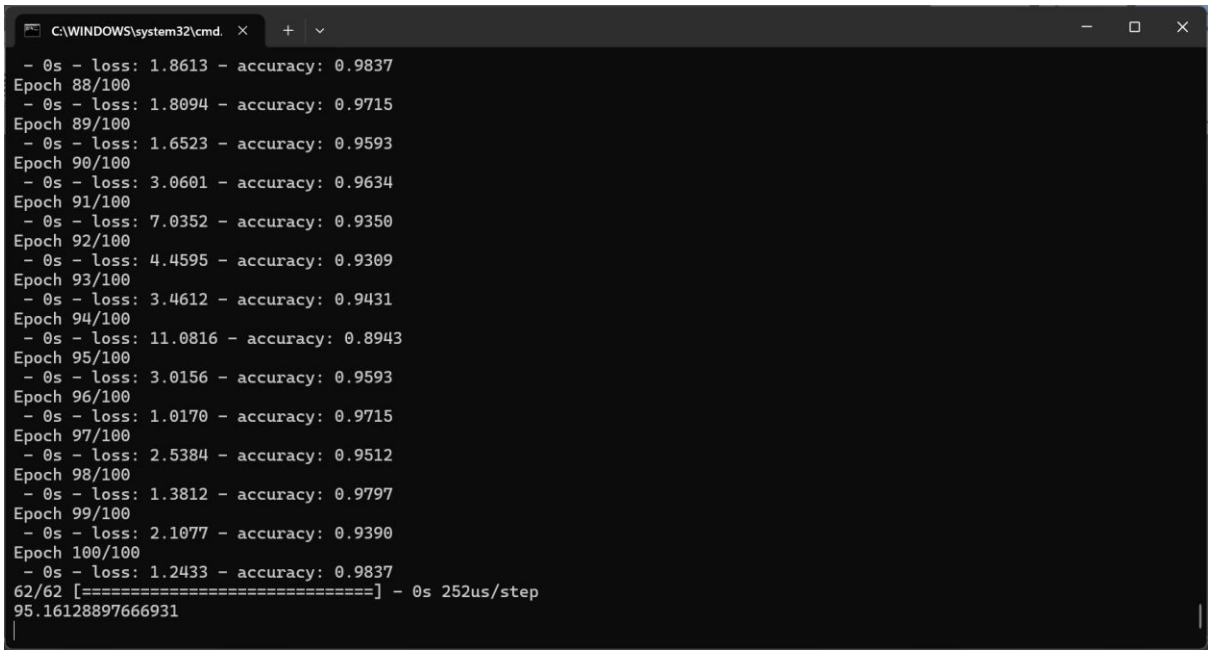
In below screen we can see SVM start training and prediction and we can see accuracy



now click on 'Run RF Algorithm' button to RF algorithm

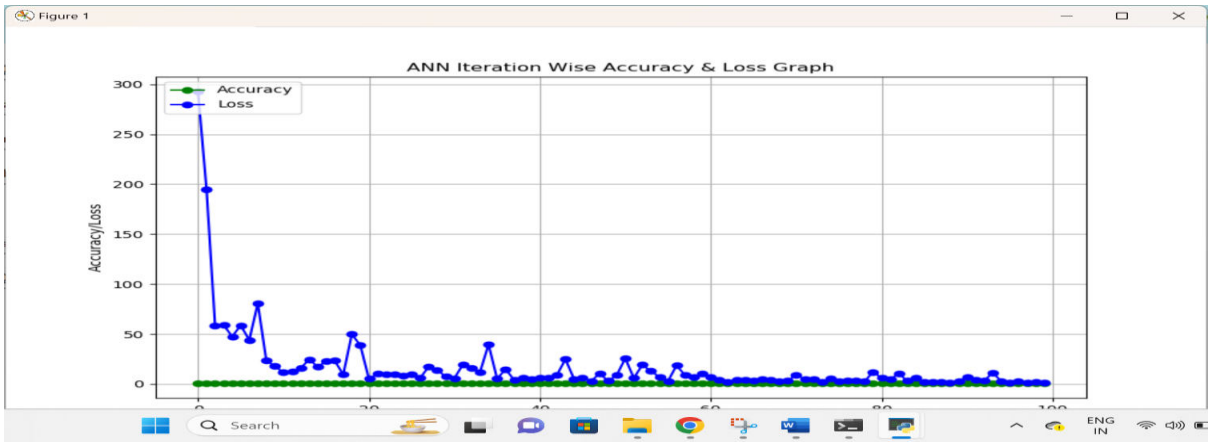
In below screen we can see RF start training and prediction and we can see accuracy





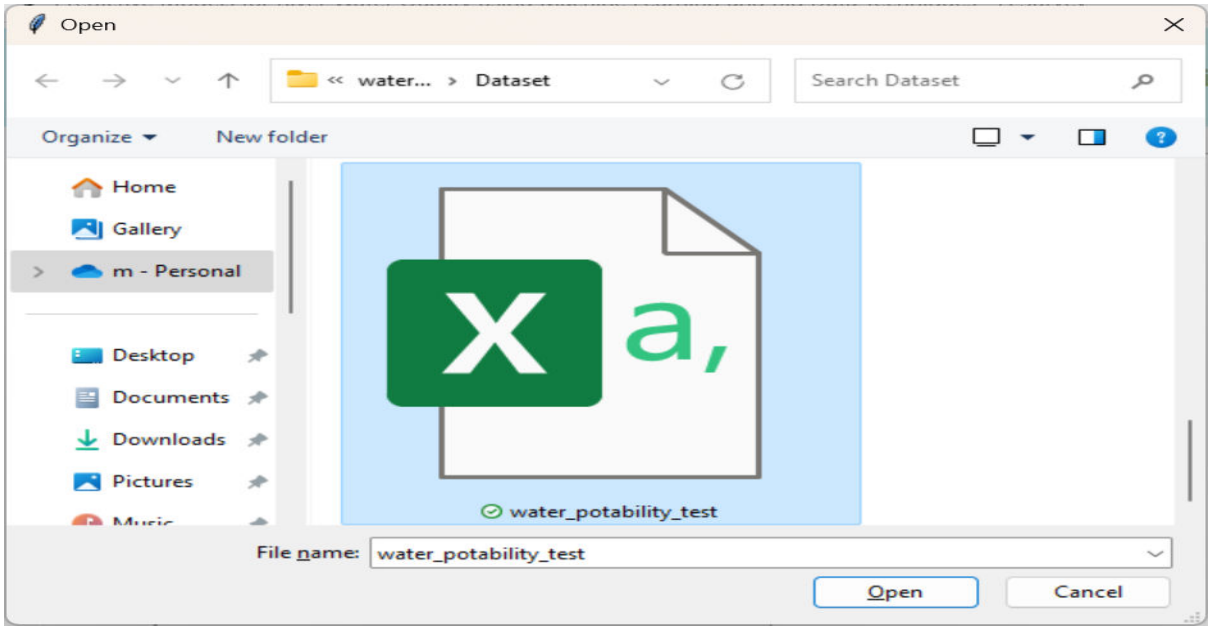
NN accuracy

'NN Accuracy & Loss Graph' button to get below graph

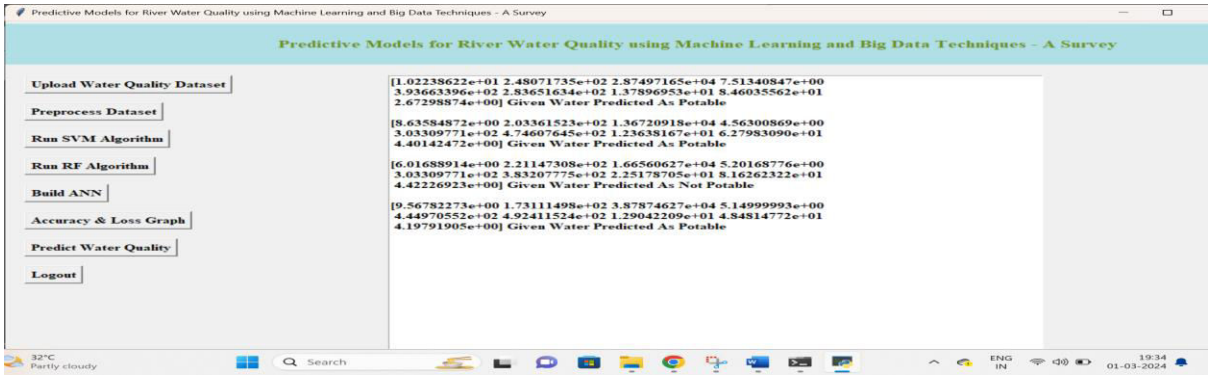


In above graph x-axis represents epoch and y-axis represents accuracy/loss value and in above graph green line represents accuracy and blue line represents loss value and loss value decrease from 7 to 0.1.

Now model is ready and now click on ‘Predict water Potability’ button to upload test data and then NN will predict below result



In above screen in square bracket, we can see uploaded test data and after square bracket we can see NN prediction result as water is potable or no



VIII. CONCLUSION

This research presents a machine learning-based system for predicting water quality and determining its potability. The proposed approach integrates multiple algorithms, including Support Vector Machine, Random Forest, and Artificial Neural Networks, to improve prediction accuracy and reliability. By leveraging data-driven techniques, the system effectively analyzes physicochemical parameters and classifies water samples.The implementation demonstrates that machine learning models can significantly enhance water quality prediction compared to traditional methods. The use of multiple algorithms ensures robustness, while the neural network captures complex relationships among variables. Experimental observations and recent studies confirm that advanced machine learning techniques achieve high accuracy and efficiency in water quality prediction .One of the key advantages of the proposed system is its scalability and adaptability. The system can handle large datasets and be updated with new data to improve performance. This is particularly important in dynamic environments where water quality conditions change frequently. Additionally, the graphical user interface enhances usability, making the system accessible to non-technical users.The system also provides real-time prediction capabilities, enabling faster decision-making and reducing reliance on traditional laboratory testing. This makes it suitable for applications in environmental monitoring, public health, and resource management.In conclusion, the proposed system offers an efficient, scalable, and accurate solution for water quality prediction. It demonstrates the potential of machine learning in addressing environmental challenges and improving water resource management. Future work can focus on integrating IoT-based data collection and advanced deep learning models to further enhance system performance.

REFERENCES

1. A. Helaly et al., “Advancements in Water Quality Prediction Using ML and DL,” *Cluster Computing*, 2025.
2. J. Zhang et al., “Machine Learning for Water Quality Prediction: A Review,” *JMSE*, 2024.
3. D. Campos et al., “AutoML for Water Quality Prediction,” *Scientific Reports*, 2026.
4. S. Patil et al., “Water Quality Analysis Using ML,” *JISEM*, 2025.
5. M. Nallakaruppan et al., “Explainable AI for Water Quality Prediction,” *Scientific Reports*, 2024.
6. A. Al-Mukhtar et al., “ML-Based Water Quality Index Prediction,” *Processes*, 2025.
7. M. A. Hridoy et al., “Advanced ML Models for Water Quality Classification,” *Science of the Total Environment*, 2025.
8. M. Deshpande et al., “ANN-Based Water Potability Prediction,” 2025.
9. I. Kaur et al., “Water Quality Assessment Using ML,” 2024.
10. A. Kuthe et al., “Water Quality Prediction Using ML,” 2023.
11. Y. Li et al., “ML Models for Water Quality Forecasting,” 2023.
12. M. Cardia et al., “Multivariate ML for Water Quality Estimation,” 2025.

13. S. Deshmukh et al., “Deep Learning for Water Monitoring,” 2025.
14. X. Xia et al., “Trustworthy ML for Water Quality Prediction,” 2025.
15. Recent Advances in ML-Based Environmental Monitoring Systems, 2024.