EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING ENSEMBLE LEARNING

PRESENTED BY

A.RAFFI

GUIDED BY: Dr. T. MYTHILI

INTRODUCTION

- **Water is considered as a vital resource that affects various aspects of human health and lives. Water quality analysis and prediction play a critical role as deteriorating water quality can have serious implications for public health, agriculture, and ecosystems.**

- **Traditional methods of water quality assessment often rely on the methods or algorithms that struggles with the dataset that doesn’t exhibit significant frequency effects and some have the problem of mishandling or misinterpretations to handle varied water conditions and parameters.**

- **To address these challenges, machine learning has emerged as a powerful tool for efficient water quality prediction, leveraging data-driven models to forecast the Water Quality Index (WQI) based on measurable parameters.**

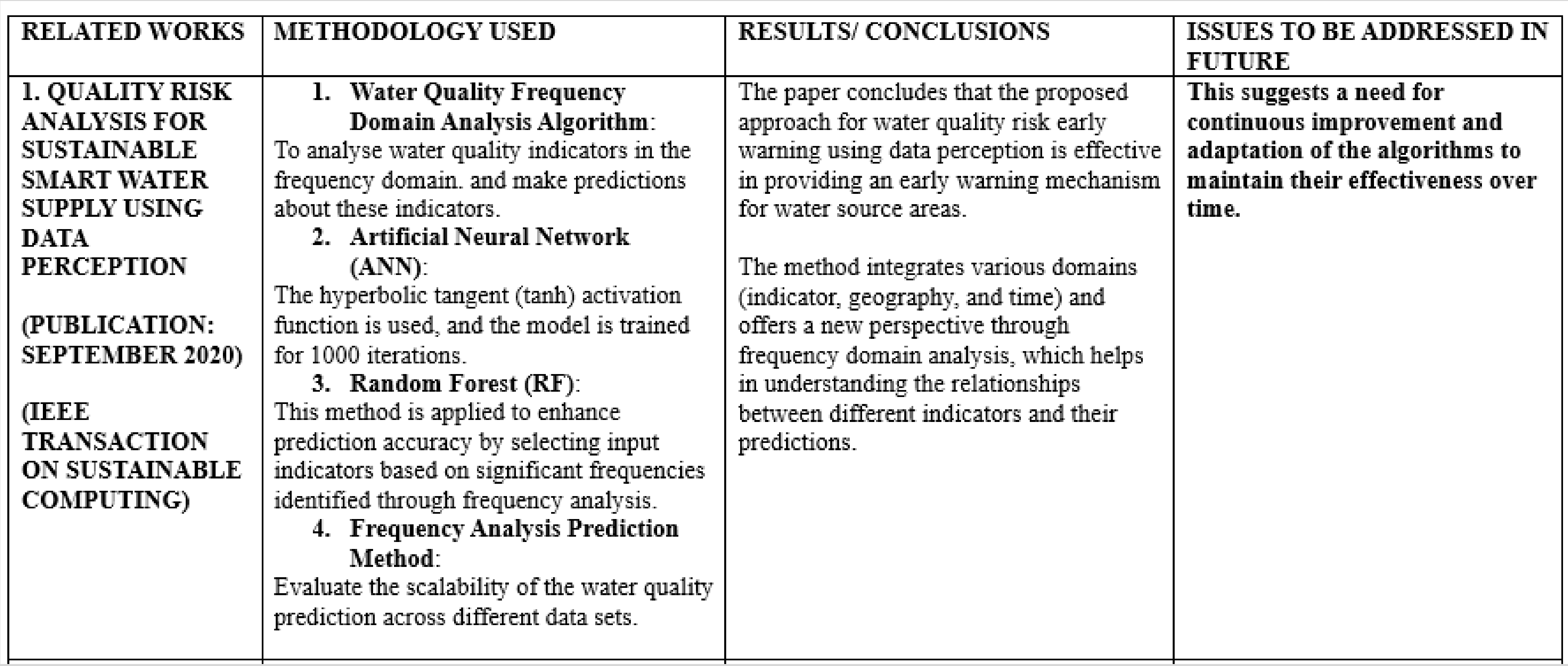
ABSTRACT

Water quality analysis and prediction play a critical role in maintaining the safety and sustainability of water resources, as deteriorating water quality can have serious implications for public health, agriculture, and ecosystems.

Traditional methods of water quality assessment often rely on the methods or algorithms that struggles with the dataset that doesn't exhibit significant frequency effects and some have the problem of mishandling or misinterpretations to handle varied water conditions and parameters.

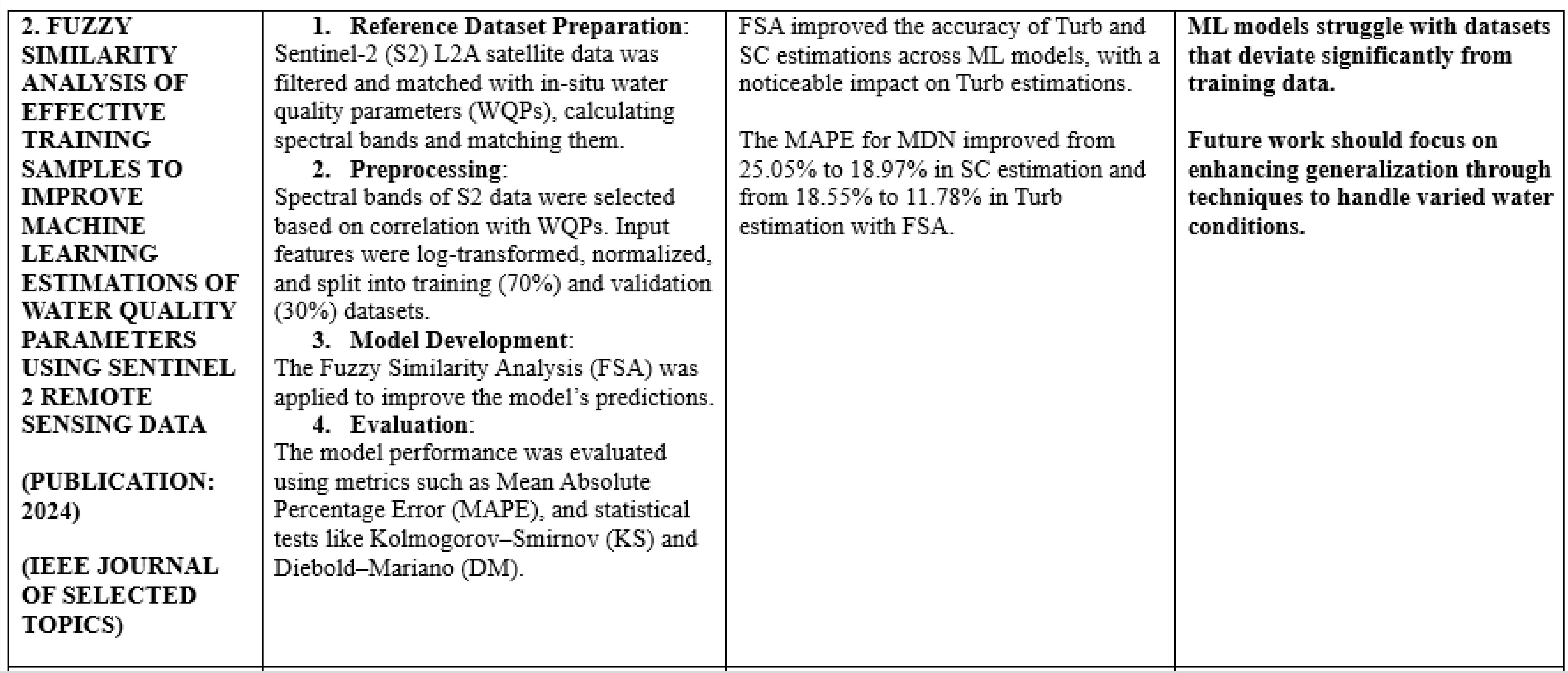
To address these challenges, this work involves data preprocessing, enhanced feature extraction methods (RFE + SHAP) to provide better interpretability of the dataset parameters and tries to analyse and work on different combination of ensemble models including Random Forest Regression, Gradient Boosting, XG Boost Regressor and Support Vector Regressor and comparative analysis of these models along with related works ensures that these ensemble model enhances water quality analysis.

LITERATURE SURVEY



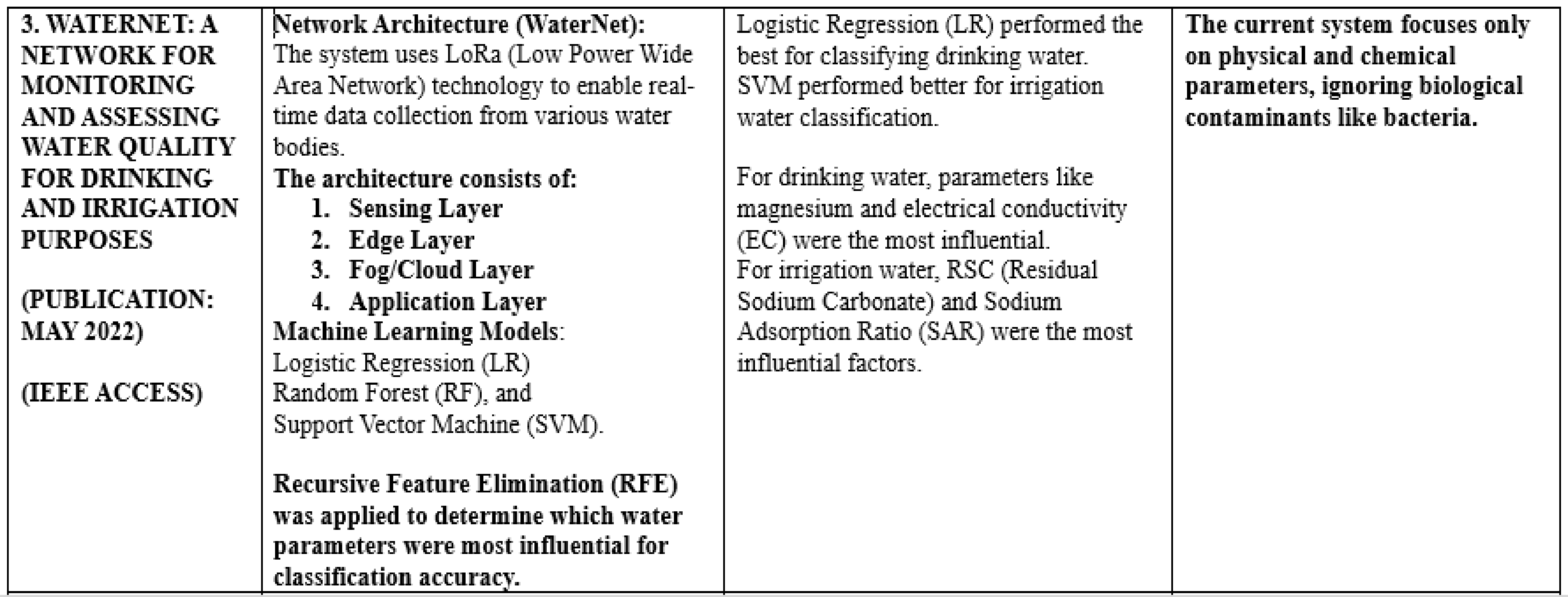
RELATED WORKS	METHODOLOGY USED	RESULTS/ CONCLUSIONS	ISSUES TO BE ADDRESSED IN FUTURE
1. QUALITY RISK ANALYSIS FOR SUSTAINABLE SMART WATER SUPPLY USING DATA PERCEPTION **(PUBLICATION: SEPTEMBER 2020)** **(IEEE TRANSACTION ON SUSTAINABLE COMPUTING)**	1. **Water Quality Frequency Domain Analysis Algorithm**: To analyse water quality indicators in the frequency domain. and make predictions about these indicators. 2. **Artificial Neural Network (ANN)**: The hyperbolic tangent (tanh) activation function is used, and the model is trained for 1000 iterations. 3. **Random Forest (RF)**: This method is applied to enhance prediction accuracy by selecting input indicators based on significant frequencies identified through frequency analysis. 4. **Frequency Analysis Prediction Method**: Evaluate the scalability of the water quality prediction across different data sets.	The paper concludes that the proposed approach for water quality risk early warning using data perception is effective in providing an early warning mechanism for water source areas. The method integrates various domains (indicator, geography, and time) and offers a new perspective through frequency domain analysis, which helps in understanding the relationships between different indicators and their predictions.	**This suggests a need for continuous improvement and adaptation of the algorithms to maintain their effectiveness over time.**

LITERATURE SURVEY



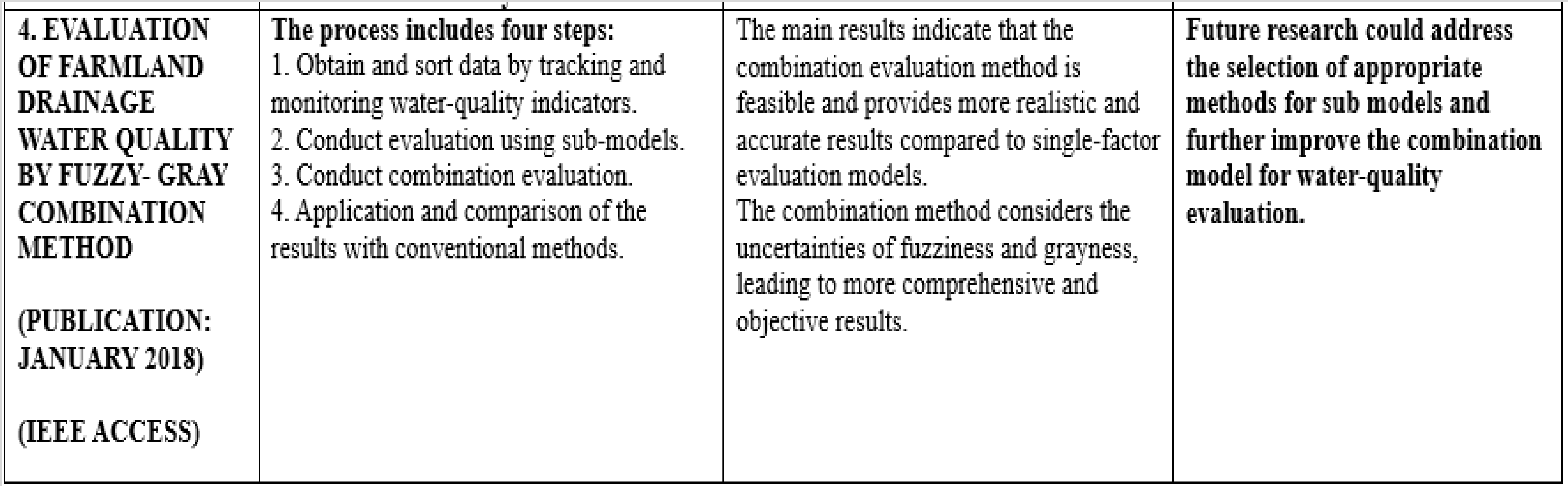
2. FUZZY SIMILARITY ANALYSIS OF EFFECTIVE TRAINING SAMPLES TO IMPROVE MACHINE LEARNING ESTIMATIONS OF WATER QUALITY PARAMETERS USING SENTINEL 2 REMOTE SENSING DATA **(PUBLICATION: 2024)** **(IEEE JOURNAL OF SELECTED TOPICS)**	1. **Reference Dataset Preparation**: Sentinel-2 (S2) L2A satellite data was filtered and matched with in-situ water quality parameters (WQPs), calculating spectral bands and matching them. 2. **Preprocessing**: Spectral bands of S2 data were selected based on correlation with WQPs. Input features were log-transformed, normalized, and split into training (70%) and validation (30%) datasets. 3. **Model Development**: The Fuzzy Similarity Analysis (FSA) was applied to improve the model's predictions. 4. **Evaluation**: The model performance was evaluated using metrics such as Mean Absolute Percentage Error (MAPE), and statistical tests like Kolmogorov–Smirnov (KS) and Diebold–Mariano (DM).	FSA improved the accuracy of Turb and SC estimations across ML models, with a noticeable impact on Turb estimations. The MAPE for MDN improved from 25.05% to 18.97% in SC estimation and from 18.55% to 11.78% in Turb estimation with FSA.	**ML models struggle with datasets that deviate significantly from training data.** **Future work should focus on enhancing generalization through techniques to handle varied water conditions.**

LITERATURE SURVEY



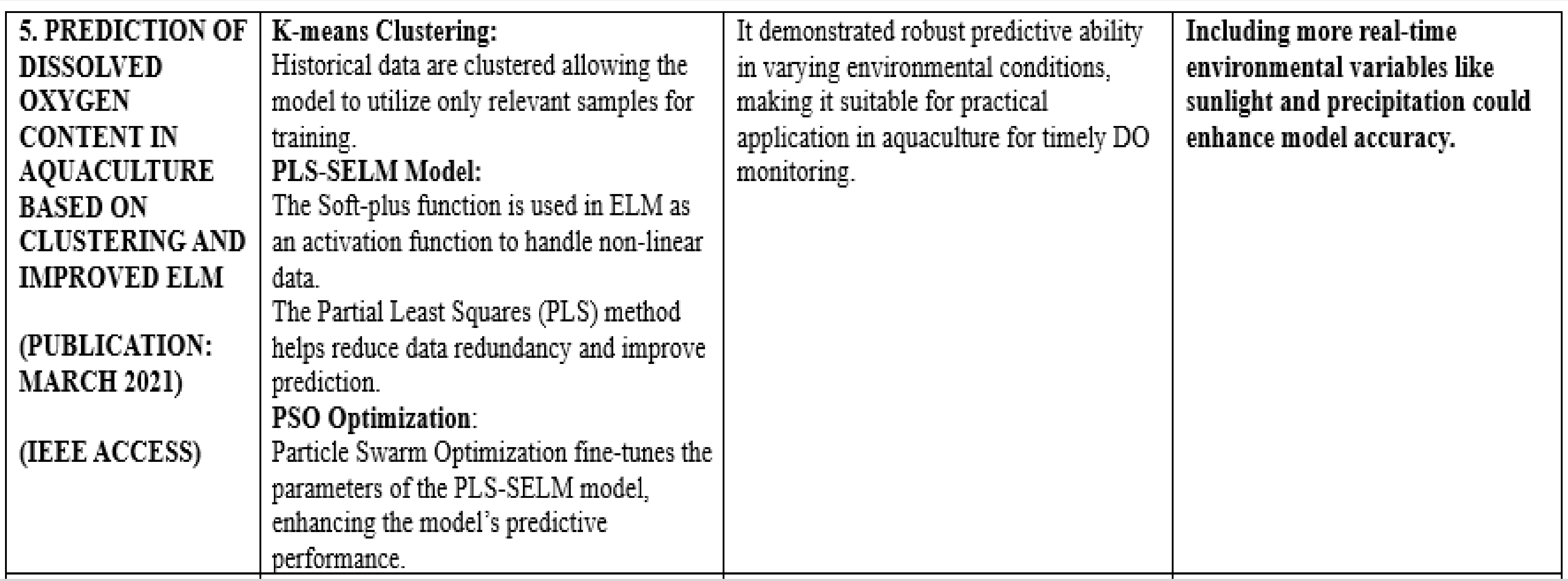
3. WATERNET: A NETWORK FOR MONITORING AND ASSESSING WATER QUALITY FOR DRINKING AND IRRIGATION PURPOSES **(PUBLICATION: MAY 2022)** **(IEEE ACCESS)**	**Network Architecture (WaterNet):** The system uses LoRa (Low Power Wide Area Network) technology to enable real-time data collection from various water bodies. **The architecture consists of:** **1. Sensing Layer** **2. Edge Layer** **3. Fog/Cloud Layer** **4. Application Layer** **Machine Learning Models**: Logistic Regression (LR) Random Forest (RF), and Support Vector Machine (SVM). **Recursive Feature Elimination (RFE) was applied to determine which water parameters were most influential for classification accuracy.**	Logistic Regression (LR) performed the best for classifying drinking water. SVM performed better for irrigation water classification. For drinking water, parameters like magnesium and electrical conductivity (EC) were the most influential. For irrigation water, RSC (Residual Sodium Carbonate) and Sodium Adsorption Ratio (SAR) were the most influential factors.	**The current system focuses only on physical and chemical parameters, ignoring biological contaminants like bacteria.**

LITERATURE SURVEY



4. EVALUATION OF FARMLAND DRAINAGE WATER QUALITY BY FUZZY- GRAY COMBINATION METHOD **(PUBLICATION: JANUARY 2018)** **(IEEE ACCESS)**	**The process includes four steps:** 1. Obtain and sort data by tracking and monitoring water-quality indicators. 2. Conduct evaluation using sub-models. 3. Conduct combination evaluation. 4. Application and comparison of the results with conventional methods.	The main results indicate that the combination evaluation method is feasible and provides more realistic and accurate results compared to single-factor evaluation models. The combination method considers the uncertainties of fuzziness and grayness, leading to more comprehensive and objective results.	**Future research could address the selection of appropriate methods for sub models and further improve the combination model for water-quality evaluation.**

LITERATURE SURVEY



5. PREDICTION OF DISSOLVED OXYGEN CONTENT IN AQUACULTURE BASED ON CLUSTERING AND IMPROVED ELM **(PUBLICATION: MARCH 2021)** **(IEEE ACCESS)**	**K-means Clustering:** Historical data are clustered allowing the model to utilize only relevant samples for training. **PLS-SELM Model:** The Soft-plus function is used in ELM as an activation function to handle non-linear data. The Partial Least Squares (PLS) method helps reduce data redundancy and improve prediction. **PSO Optimization**: Particle Swarm Optimization fine-tunes the parameters of the PLS-SELM model, enhancing the model's predictive performance.	It demonstrated robust predictive ability in varying environmental conditions, making it suitable for practical application in aquaculture for timely DO monitoring.	**Including more real-time environmental variables like sunlight and precipitation could enhance model accuracy.**

OBJECTIVES

1. To develop a machine learning model to predict the Water Quality Index (WQI) using key water quality parameters.

2. To enhance model accuracy and robustness through advanced feature extraction methods.

3. To employ an ensemble approach to leverage the strengths of multiple models for improved prediction.

PROPOSED SYSTEM

- The proposed work for water quality analysis and prediction offers a range of powerful features tailored to enhance predictive accuracy and usability. First, it integrates enhanced feature extraction techniques like RFE and SHAP which capture key relationships and ensure model interpretability.

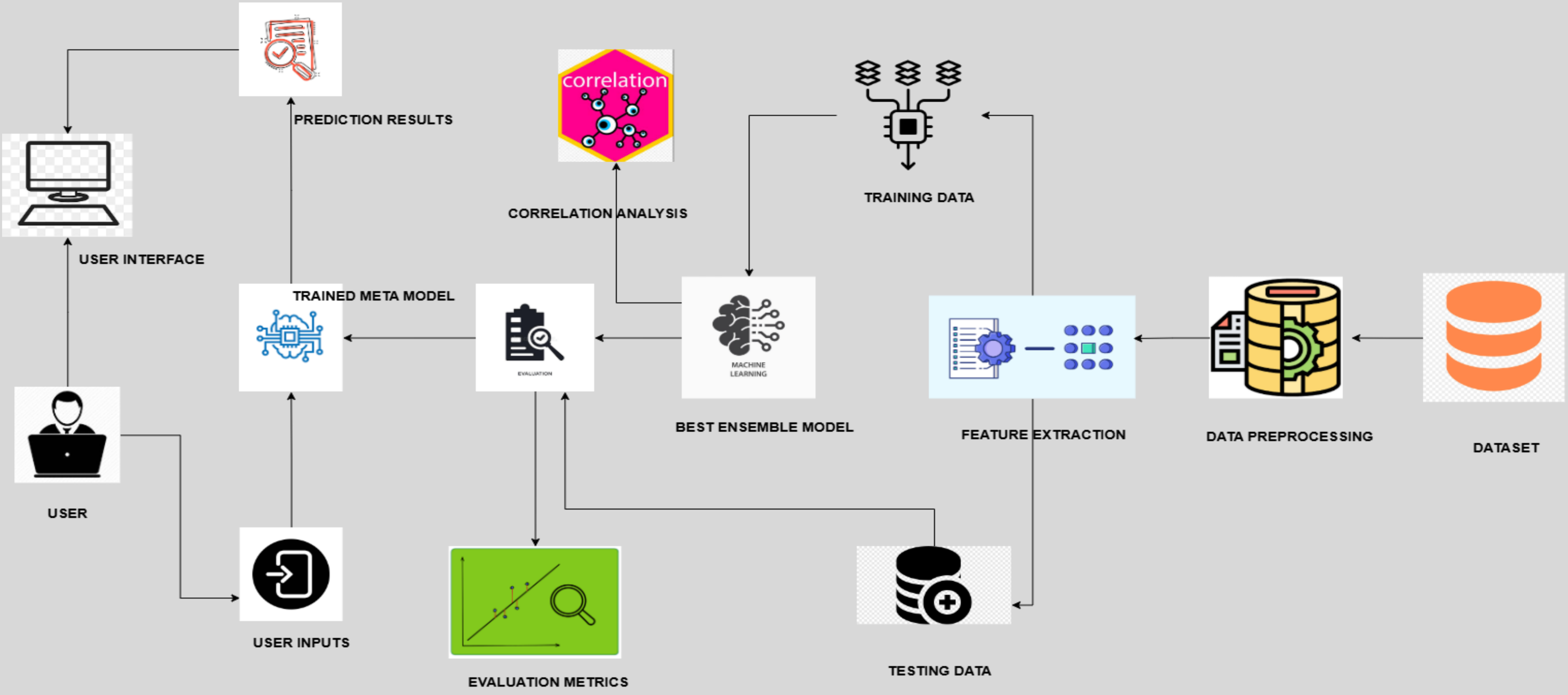
- The system employs a stacked ensemble model combining various combinations of Random Forest, Gradient Boosting, XG Boost, Support Vector Regression and Neural Networks, leveraging the strengths of each algorithm to improve prediction robustness and accuracy.

- Comparative analysis is made based on the evaluation metrics like MAE, MSE, R2 score and AUC for each of these ensemble models with the base models and the related works involved.

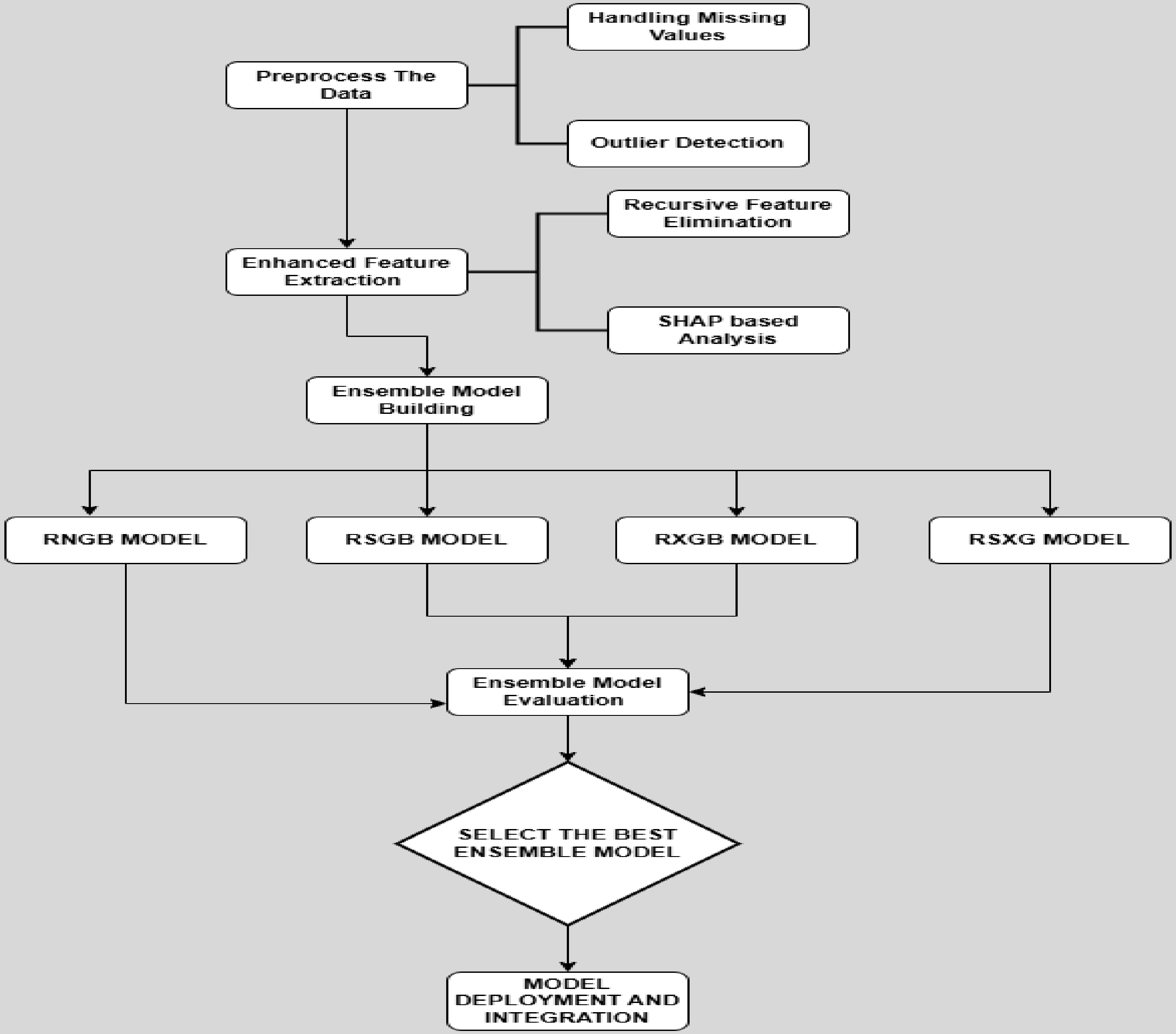
- The ensemble model with superior performance was used in our proposed work for deployment.

- Through a user-friendly interface, stakeholders can easily access insights and alerts, empowering proactive water resource management and environmental protection.

SYSTEM ARCHITECTURE



FLOWCHART



Preprocess The Data

Handling Missing Values

Outlier Detection

Enhanced Feature Extraction

Recursive Feature Elimination

SHAP based Analysis

Ensemble Model Building

RNGB MODEL

RSGB MODEL

RXGB MODEL

RSXG MODEL

Ensemble Model Evaluation

SELECT THE BEST ENSEMBLE MODEL

MODEL DEPLOYMENT AND INTEGRATION

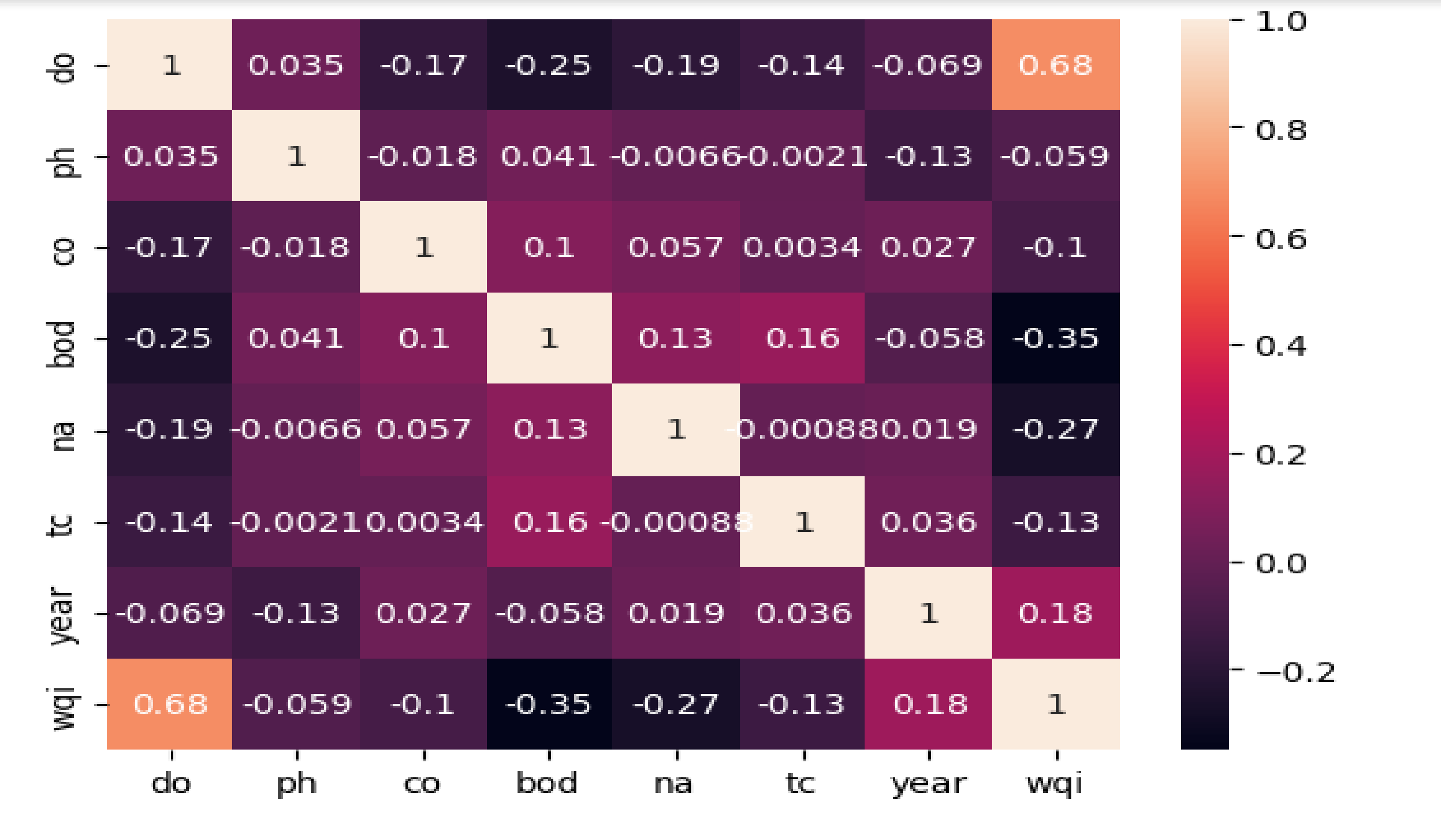
METHODOLOGY

1. DATASET COLLECTION
2. DATA PREPROCESSING
3. ENHANCED FEATURE EXTRACTION METHODS
4. ENSEMBLE MODEL BUILDING AND EVALUATION
5. DEPLOYMENT PHASE

ATTRIBUTES OF FEATURE SET

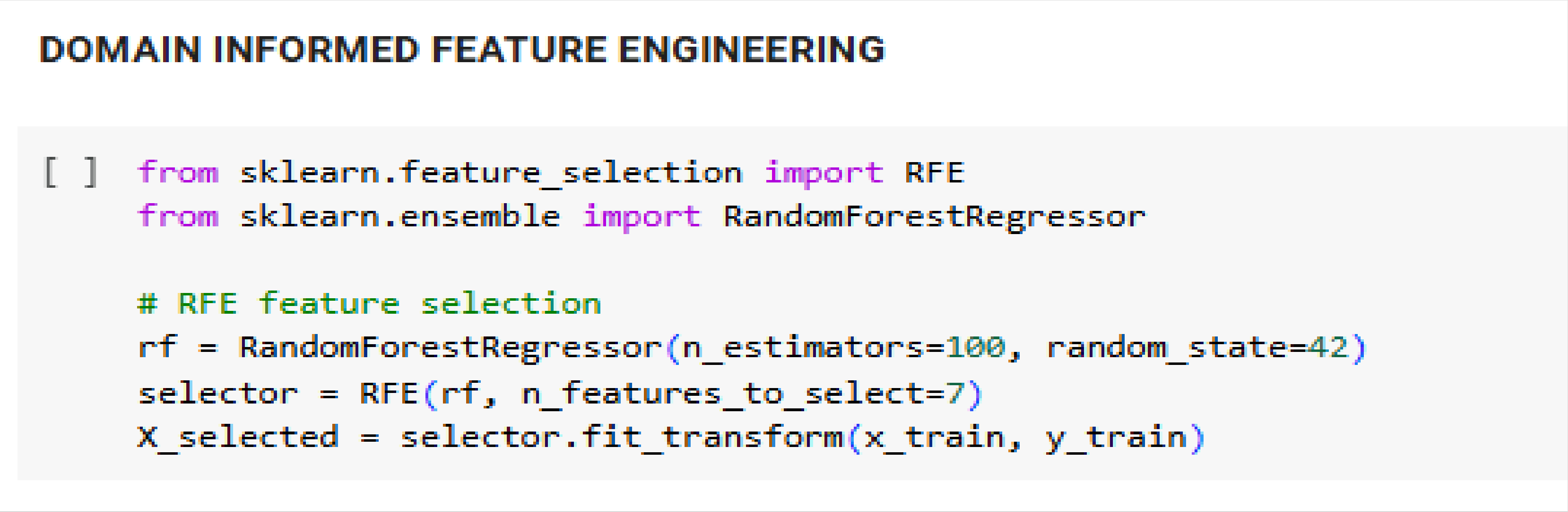
S. No	Features	Description	Significance
1	Temperature	Influences chemical and biological process involved in water.	Affects dissolved oxygen levels, biological activity and chemical reaction rates.
2	Dissolved Oxygen (D.O.) (mg/L)	Amount of oxygen dissolved in water.	Low levels indicate pollution leading to hypoxic conditions.
3	pH	Indicates the acidity or alkalinity of water.	Extreme values can influence chemical solubility, impacting metal toxicity.
4	Conductivity (μmhos/cm)	Measures the water's ability to conduct electricity, linked to dissolved ions.	High values suggest excessive dissolved salts, indicating pollution from agricultural or industrial runoff.
5	B.O.D. (mg/L)	Amount of oxygen required by microorganisms to decompose organic matter.	High B.O.D. suggests heavy organic pollution, leading to oxygen depletion and harming aquatic organisms.
6	NITRATENAN N+ NITRITENANN (mg/L)	Measures nitrogen-based compounds in water.	High levels cause eutrophication, leading to algal blooms, oxygen depletion.
7	FECAL COLIFORM (MPN/100ml)	Indicator of fecal contamination from human/animal waste.	Presence suggests potential pathogens, indicating waterborne disease risk.
8	TOTAL COLIFORM (MPN/100ml) Mean	Measures overall bacterial contamination.	High levels indicate poor sanitation, increasing the risk of waterborne infections.

DATA PREPROCESSING (CORRELATION MATRIX)



	do	ph	co	bod	na	tc	year	wqi
do	1	0.035	-0.17	-0.25	-0.19	-0.14	-0.069	0.68
ph	0.035	1	-0.018	0.041	-0.0066	-0.0021	-0.13	-0.059
co	-0.17	-0.018	1	0.1	0.057	0.0034	0.027	-0.1
bod	-0.25	0.041	0.1	1	0.13	0.16	-0.058	-0.35
na	-0.19	-0.0066	0.057	0.13	1	-0.00088	0.019	-0.27
tc	-0.14	-0.0021	0.0034	0.16	-0.00088	1	0.036	-0.13
year	-0.069	-0.13	0.027	-0.058	0.019	0.036	1	0.18
wqi	0.68	-0.059	-0.1	-0.35	-0.27	-0.13	0.18	1

ENHANCED FEATURE EXTRACTION (RFE)

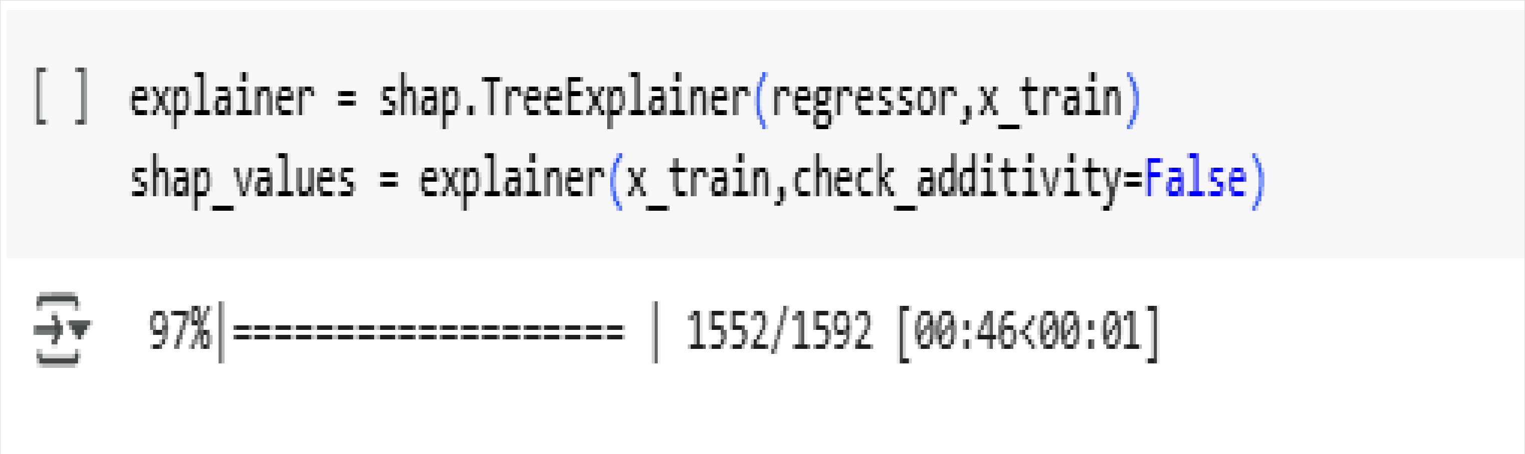


DOMAIN INFORMED FEATURE ENGINEERING

```
from sklearn.feature_selection import RFE
from sklearn.ensemble import RandomForestRegressor

# RFE feature selection
rf = RandomForestRegressor(n_estimators=100, random_state=42)
selector = RFE(rf, n_features_to_select=7)
X_selected = selector.fit_transform(x_train, y_train)
```

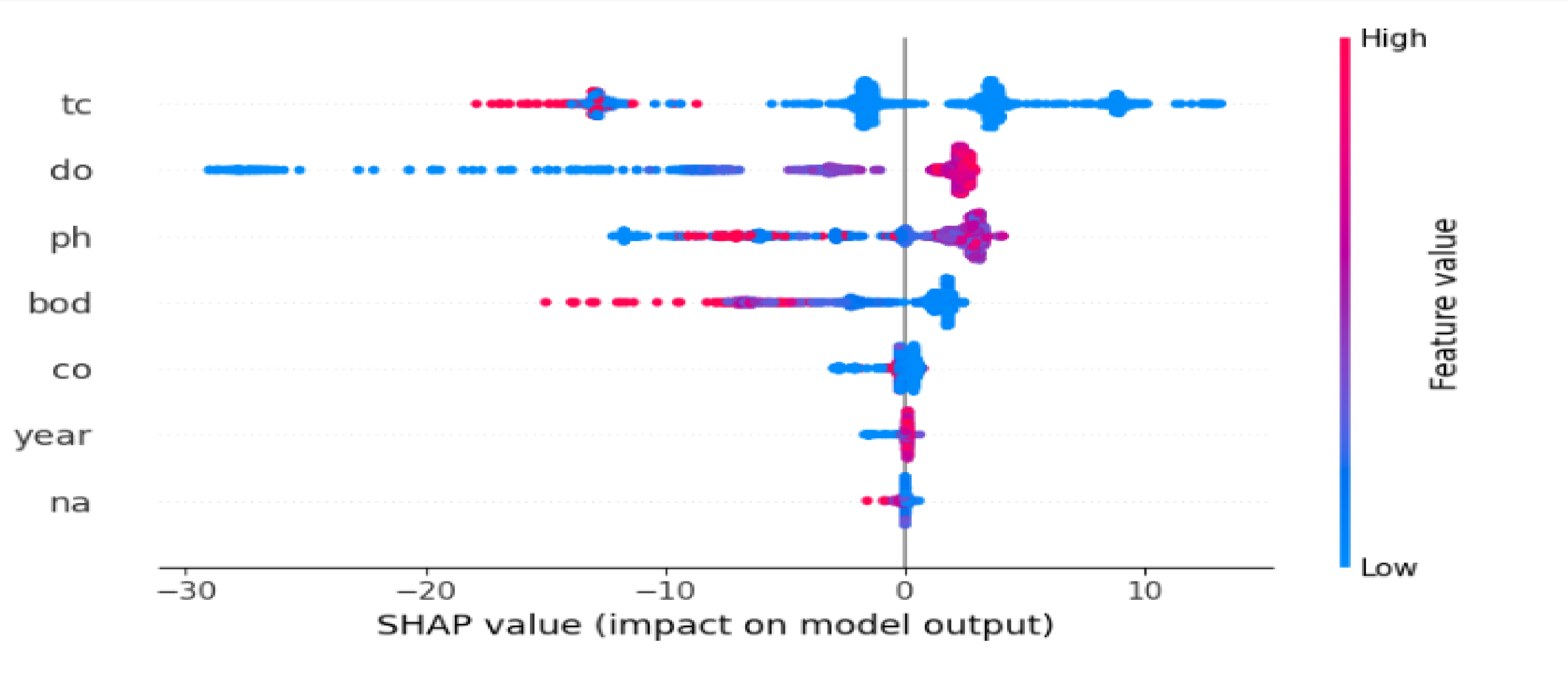
ENHANCED FEATURE EXTRACTION (SHAP BASED ANALYSIS)



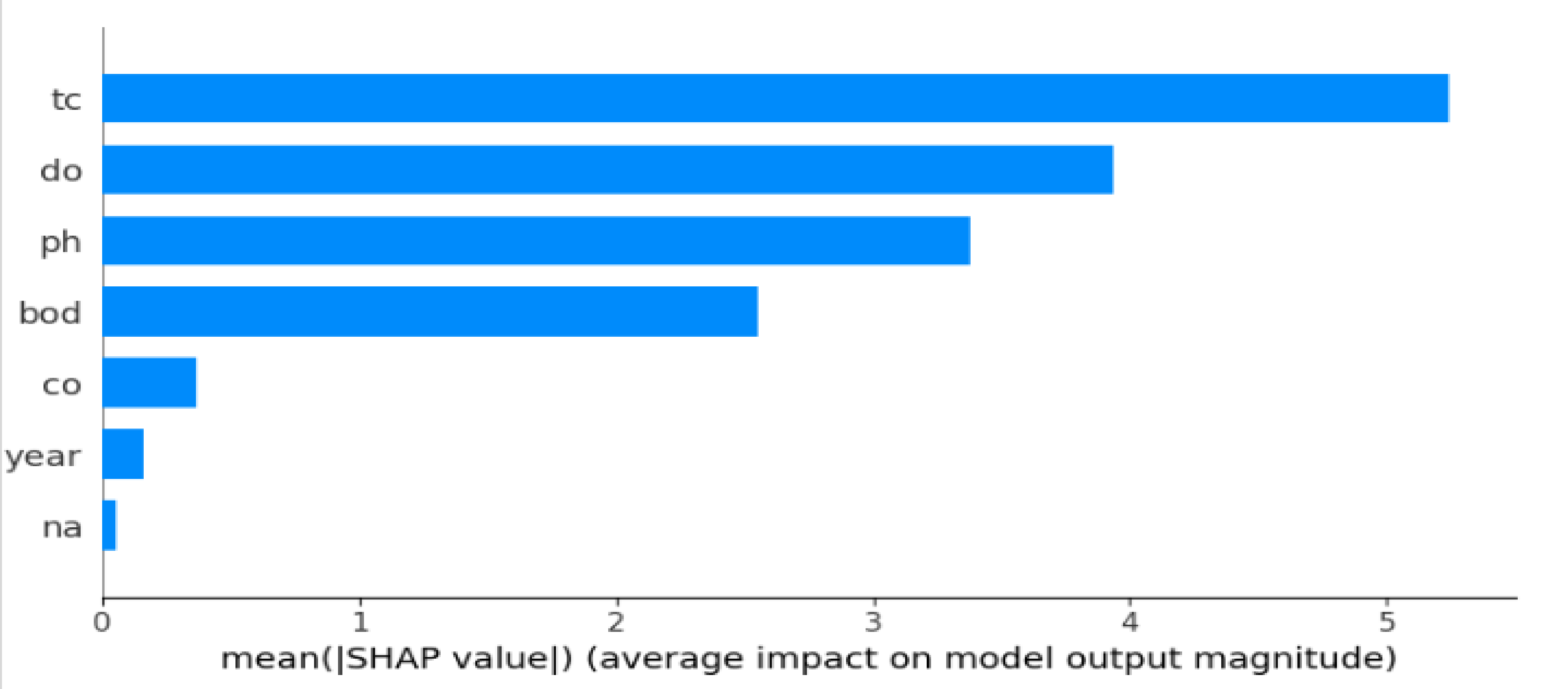
```
[ ] explainer = shap.TreeExplainer(regressor,x_train)
    shap_values = explainer(x_train,check_additivity=False)
```

```
97%|=================== | 1552/1592 [00:46<00:01]
```

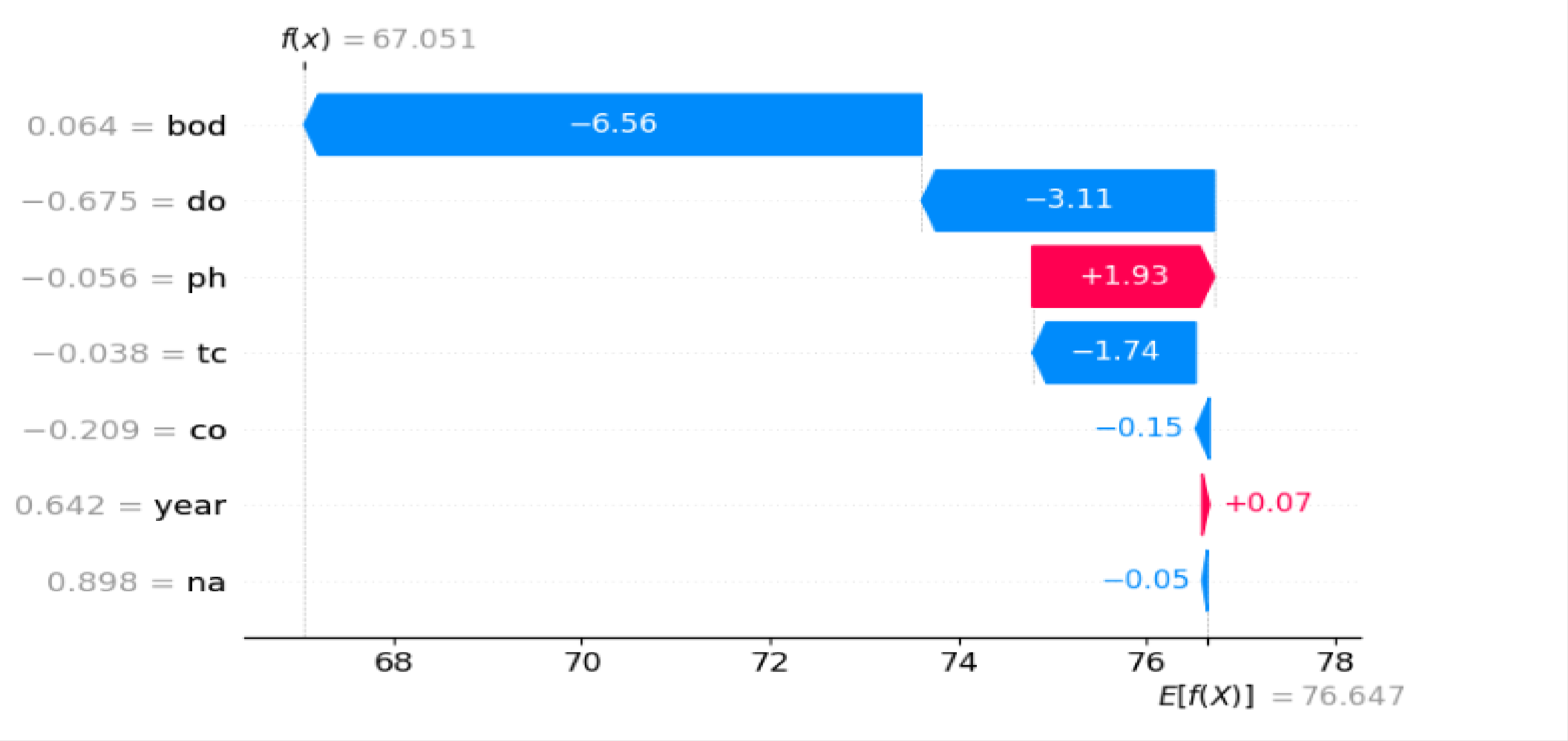
SHAP BASED FEATURE IMPORTANCE ANALYSIS



BAR PLOT


tc
do
ph
bod
co
year
na
0
1
2
3
4
5
mean(|SHAP value|) (average impact on model output magnitude)

WATERFALL PLOT



ENSEMBLE MODEL BUILDING

ENSEMBLE MODELLING COMBINATIONS

- **RNGB**

  (RANDOM FOREST REGRESSION, GRADIENT BOOSTING & NEURAL NETWORK)
- **RSGB**

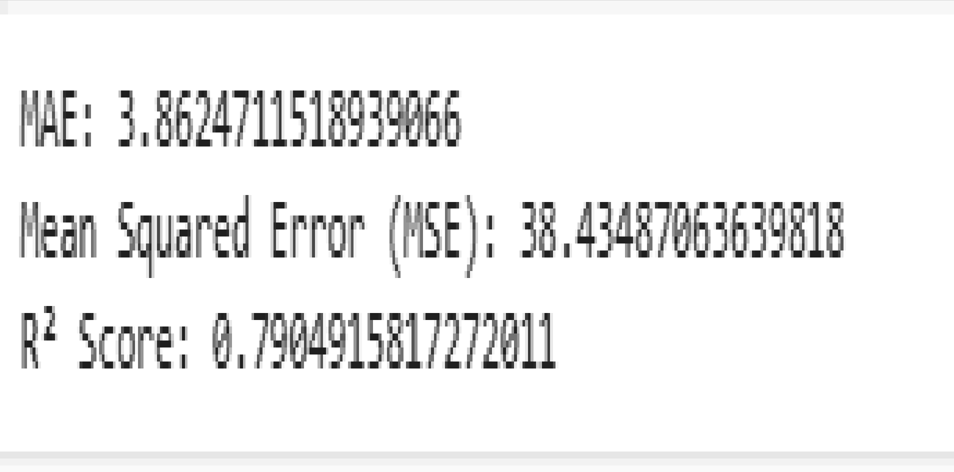
  (RANDOM FOREST REGRESSION, GRADIENT BOOSTING & SUPPORT VECTOR REGRESSOR)
- **RXGB**

  (RANDOM FOREST REGRESSION, GRADIENT BOOSTING & XG BOOST REGRESSOR)
- **RSXG**

  (RANDOM FOREST REGRESSION, XG BOOST REGRESSOR & SUPPORT VECTOR REGRESSOR)

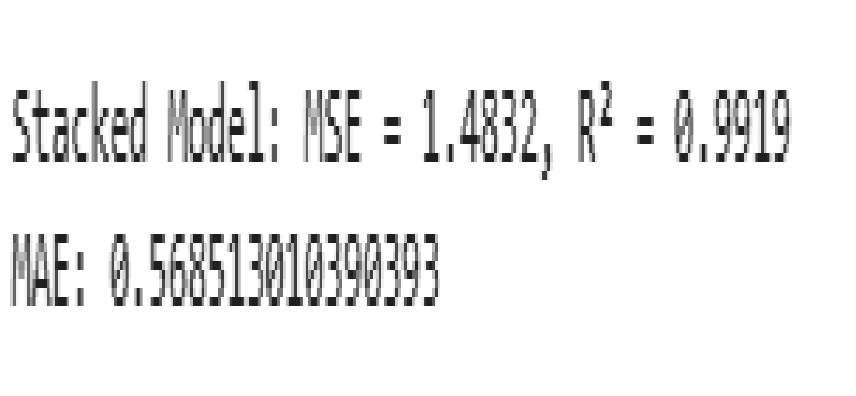
EVALUATION METRICS

- **ENSEMBLE MODEL 1: RNGB**



```
MAE: 3.862471151893906б
Mean Squared Error (MSE): 38.43487063639818
R² Score: 0.7904915817272011
```

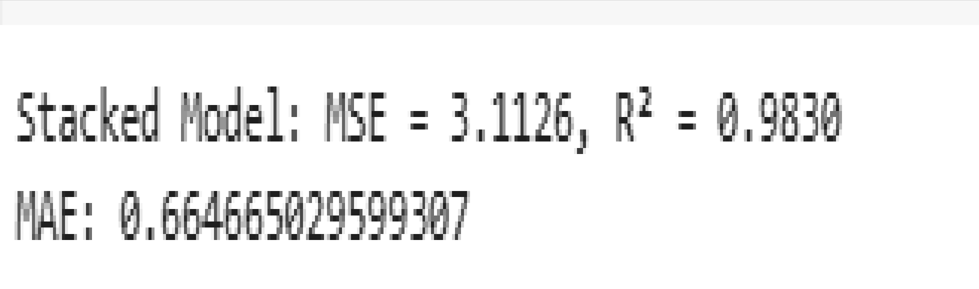
- **ENSEMBLE MODEL 2: RSGB**



```
Stacked Model: MSE = 1.4832, R² = 0.9919
MAE: 0.568513010390393
```

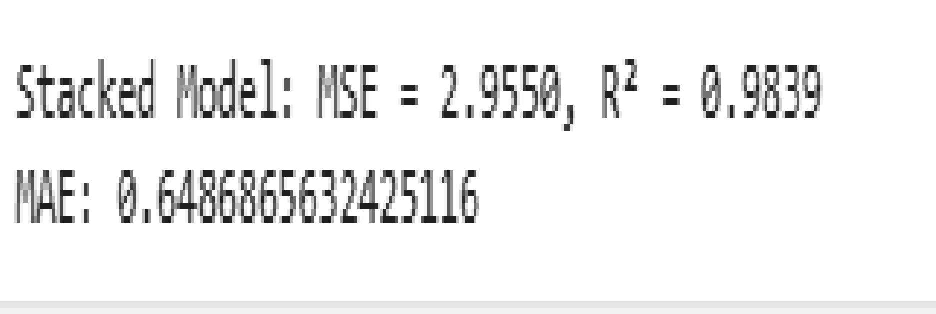
EVALUATION METRICS

- **ENSEMBLE MODEL 3: RXGB**



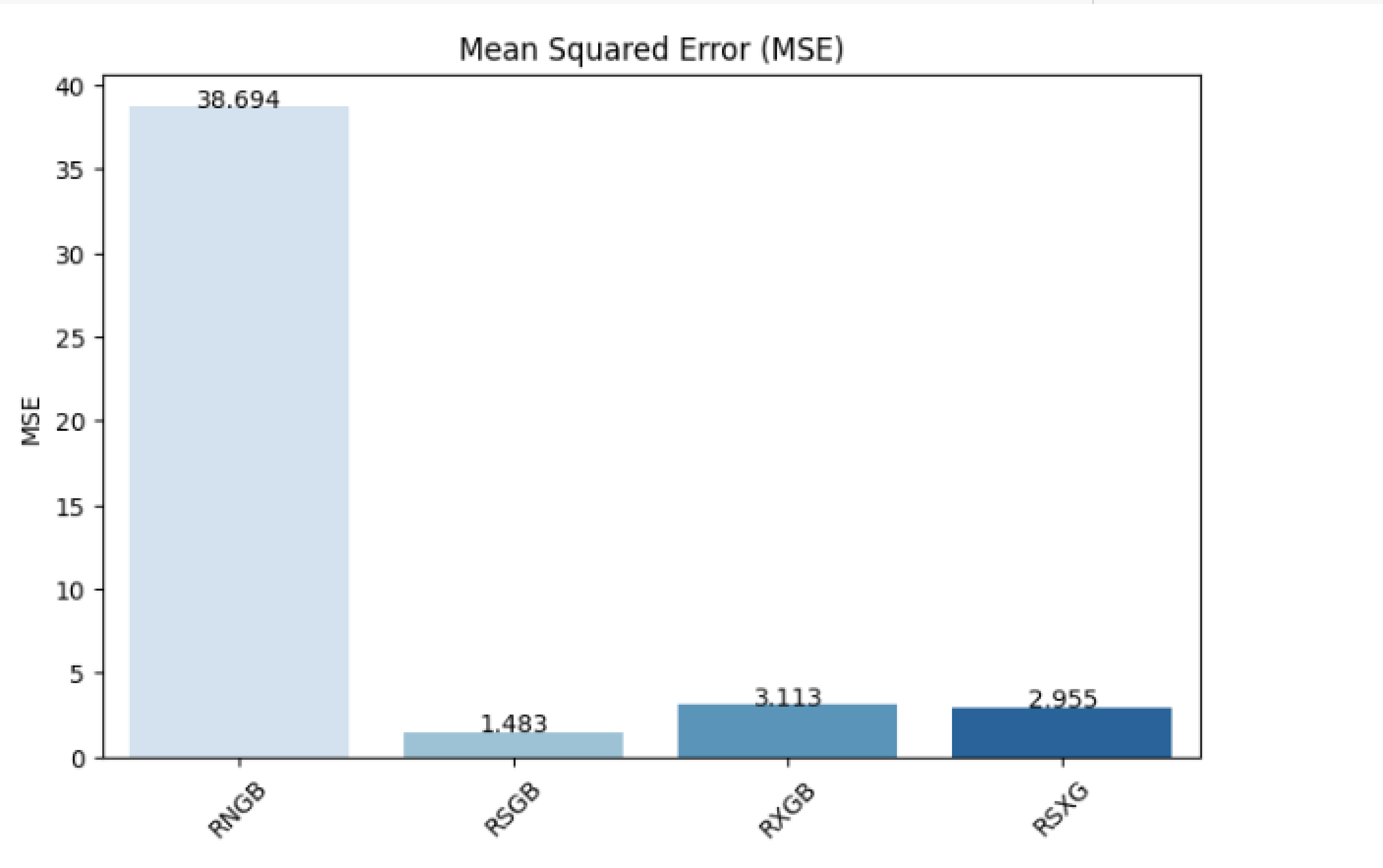
```
Stacked Model: MSE = 3.1126, R² = 0.9830
MAE: 0.664665029599307
```

- **ENSEMBLE MODEL 4: RSXG**

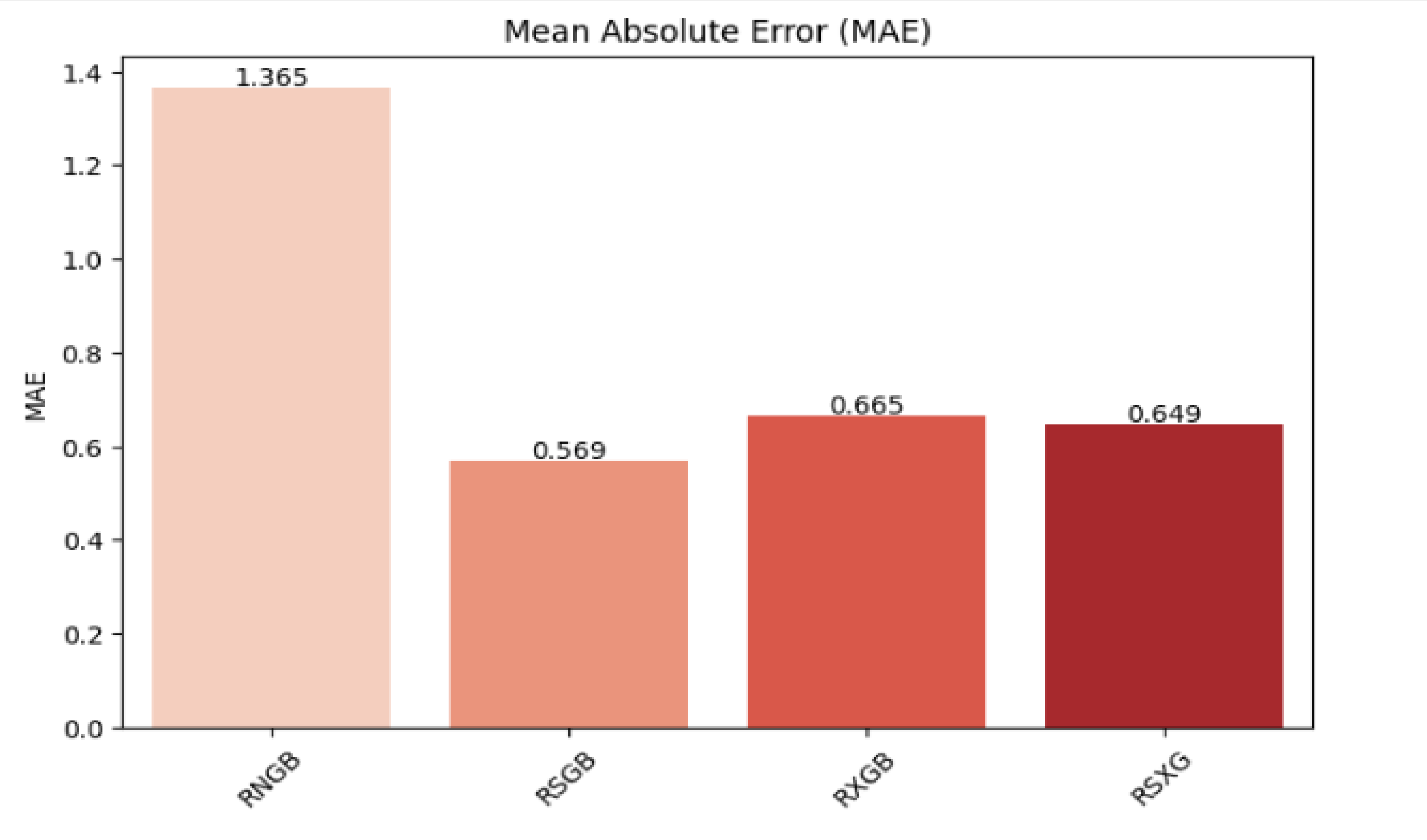


```
Stacked Model: MSE = 2.9550, R² = 0.9839
MAE: 0.6486865632425116
```

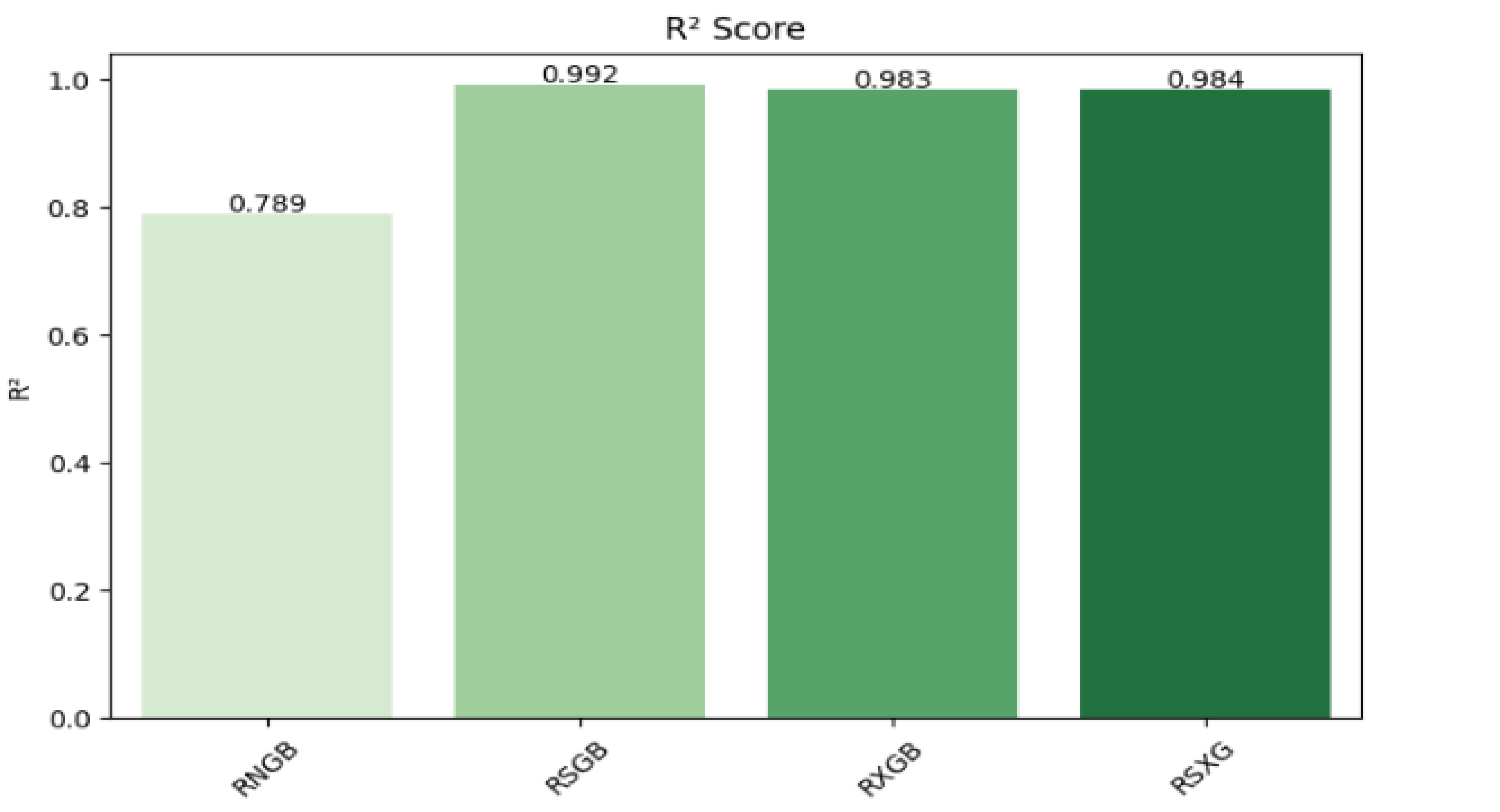
MEAN SQUARED ERROR (MSE) PLOT



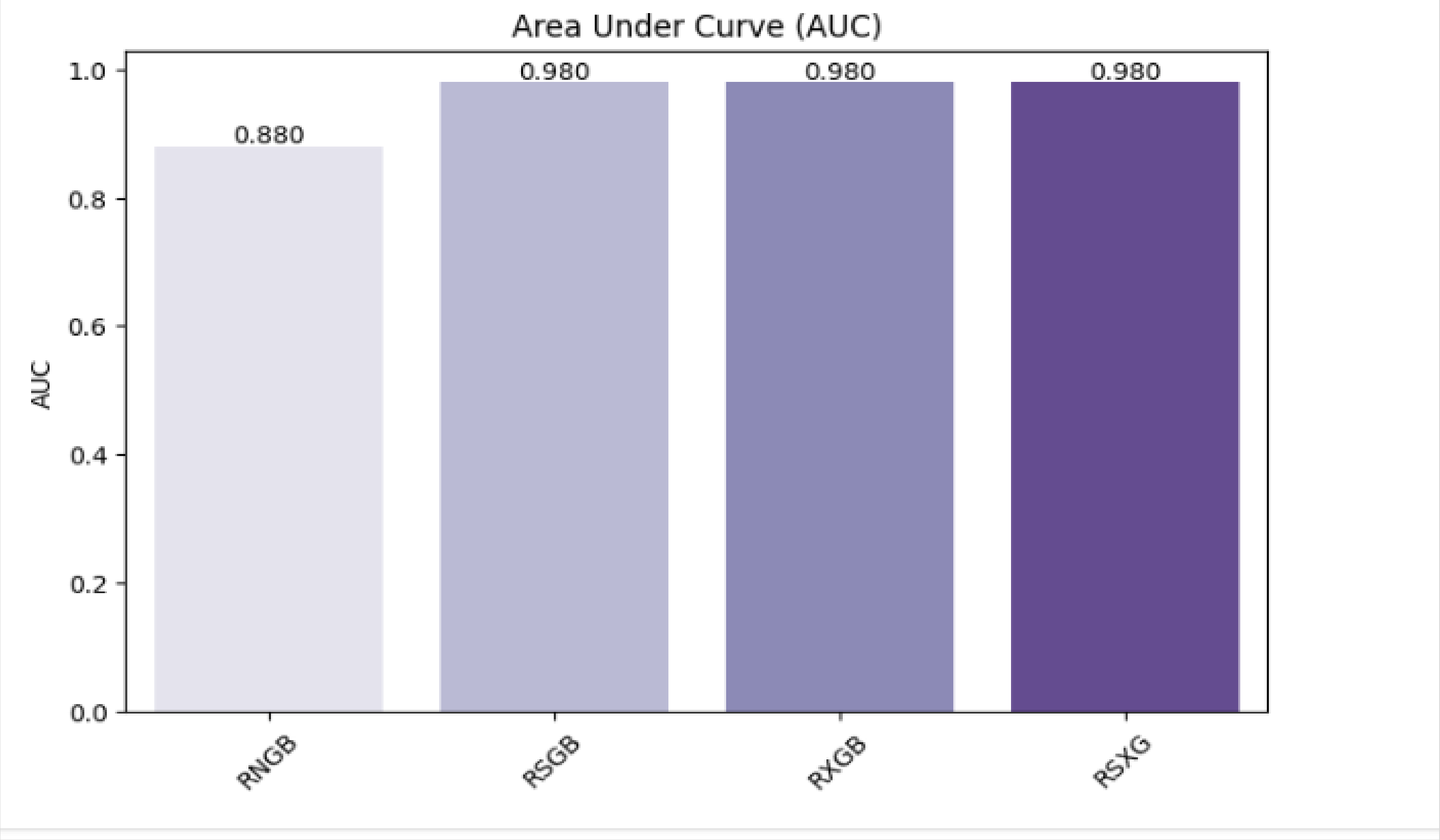
MEAN ABSOLUTE ERROR (MAE) PLOT



R2 SCORE PLOT

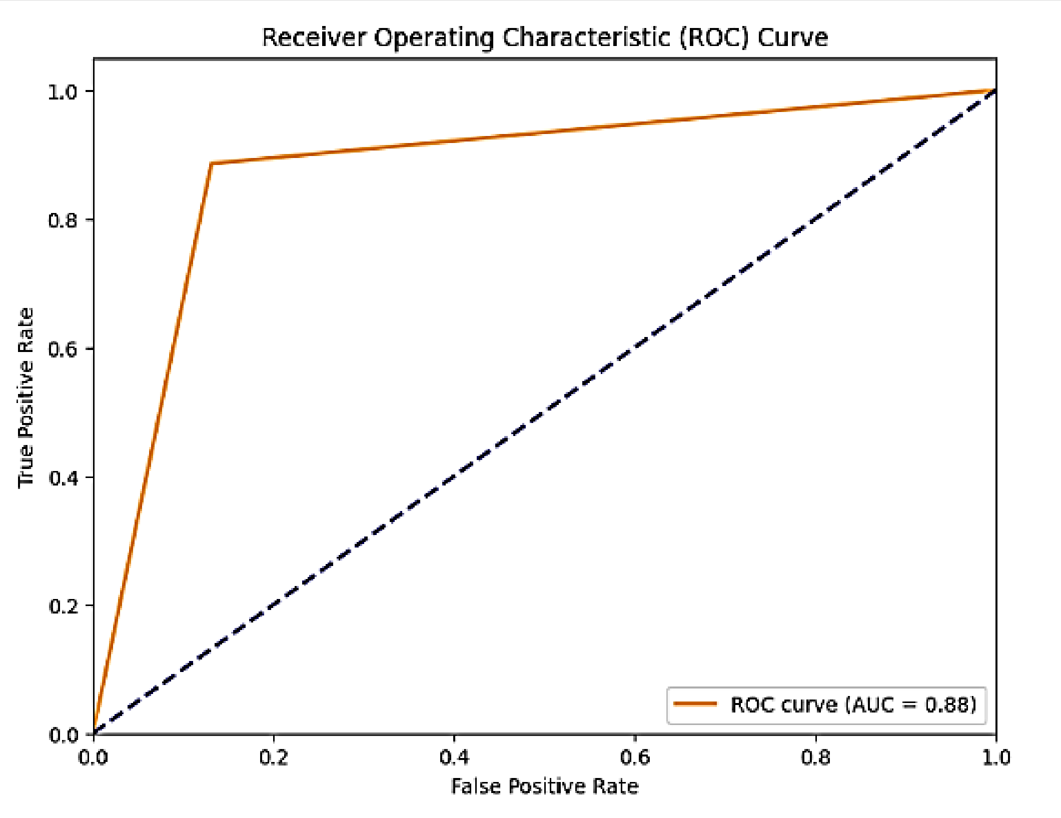


AUC PLOT

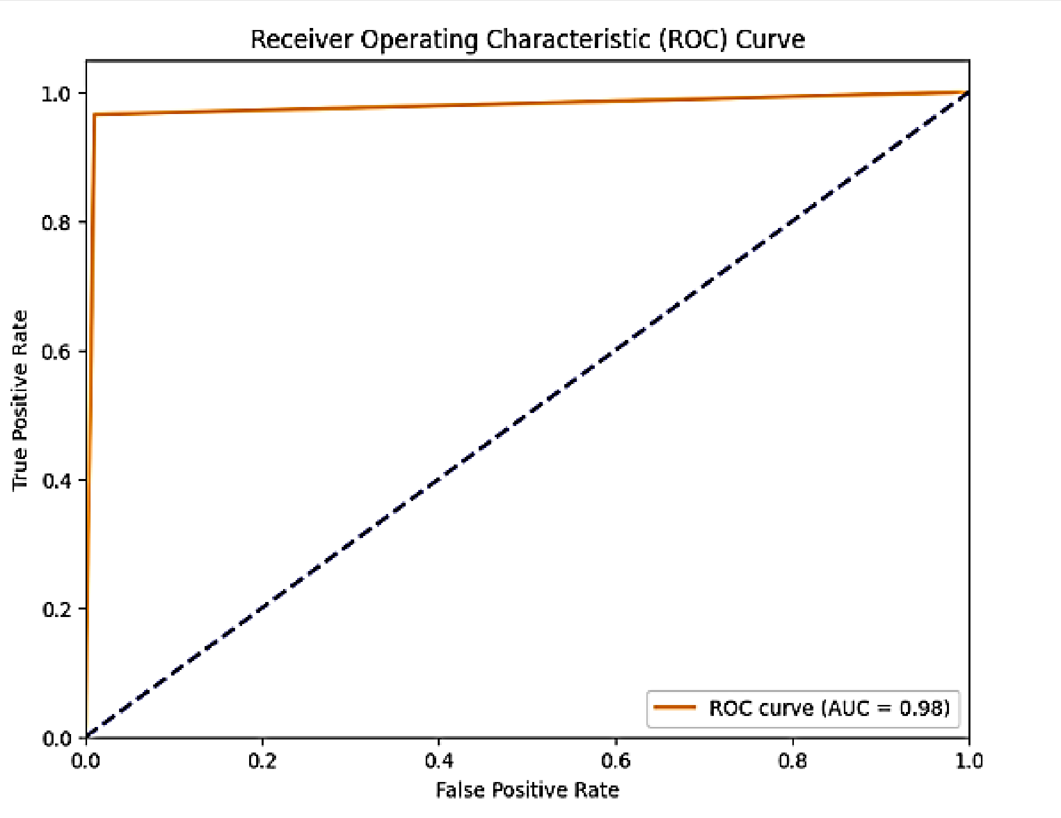

Area Under Curve (AUC)
AUC
1.0
0.8
0.6
0.4
0.2
0.0
0.880
0.980
0.980
0.980
RNGB
RSGB
RXGB
RSXG

ROC CURVE

- **ENSEMBLE MODEL 1: RNGB**

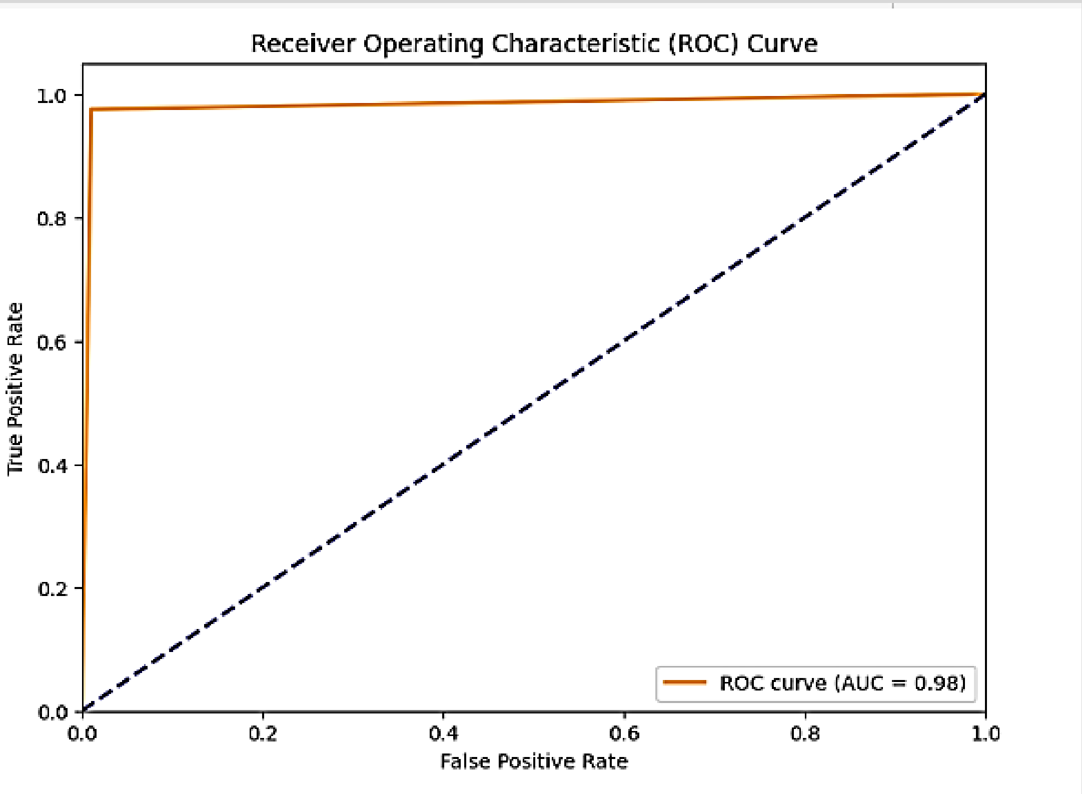


- **ENSEMBLE MODEL 2: RSGB**

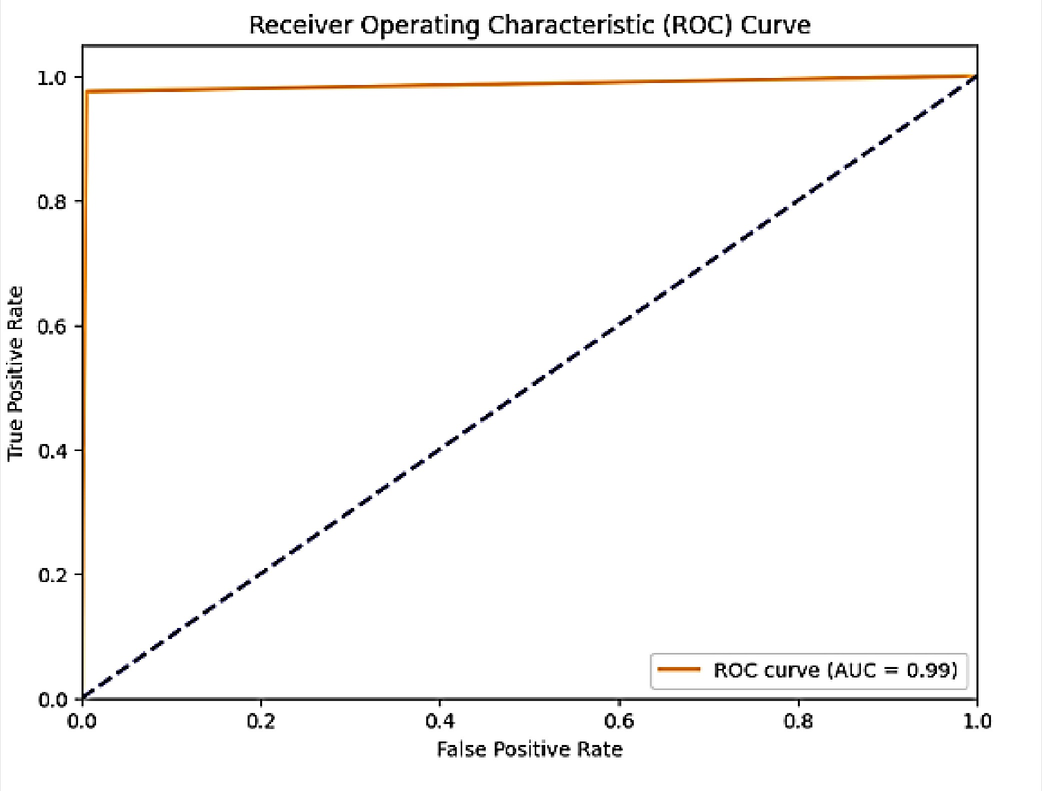


ROC CURVE

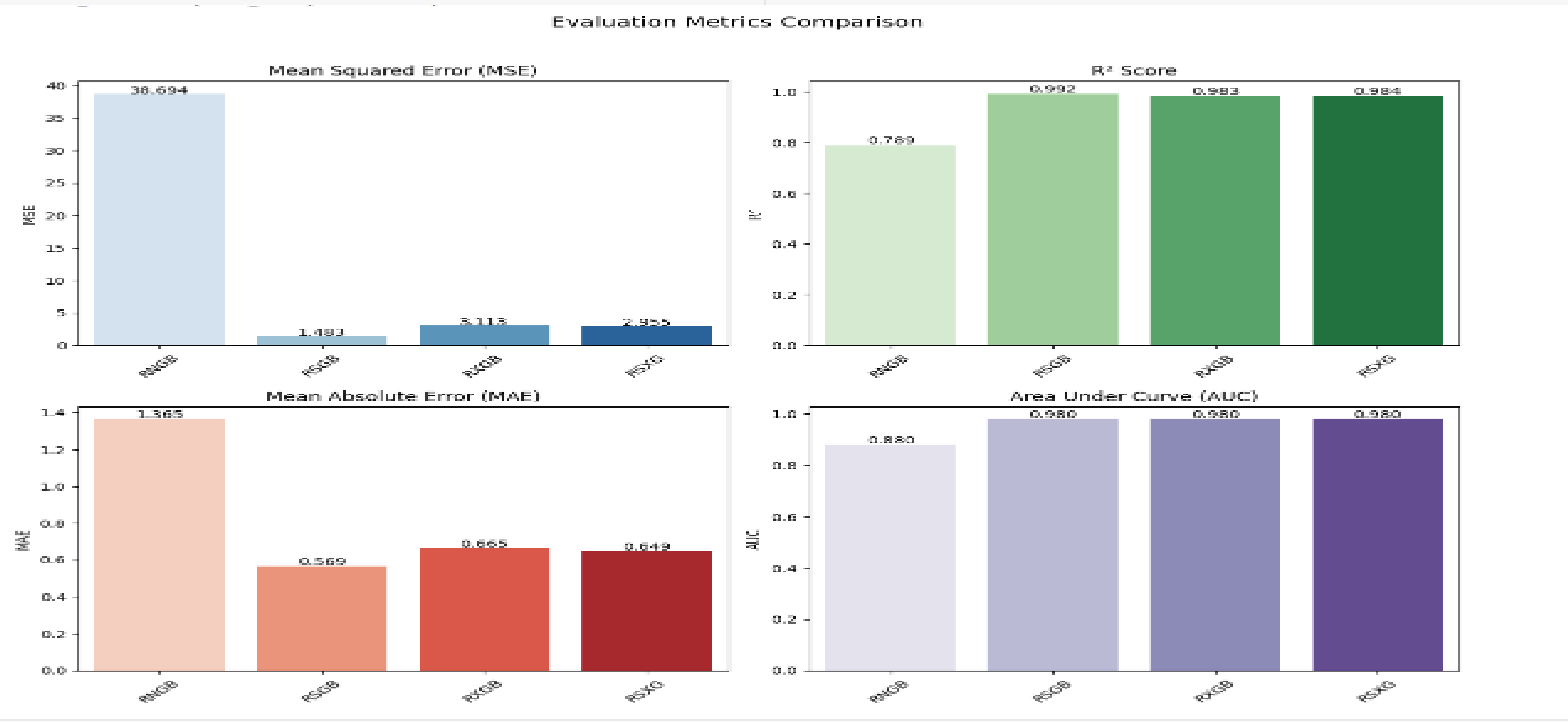
- ENSEMBLE MODEL 3: RXGB



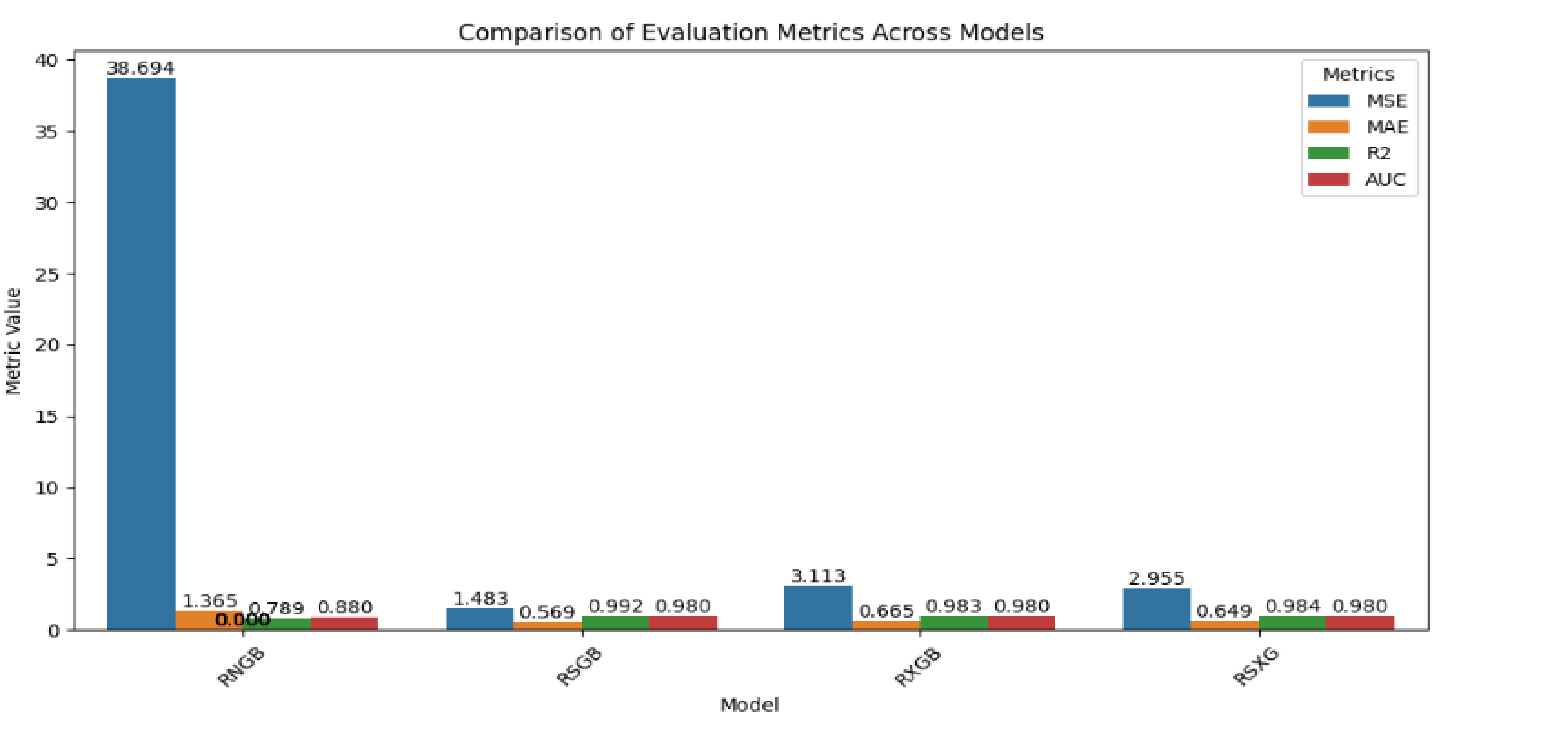
- ENSEMBLE MODEL 4: RSXG



EVALUATION METRICS



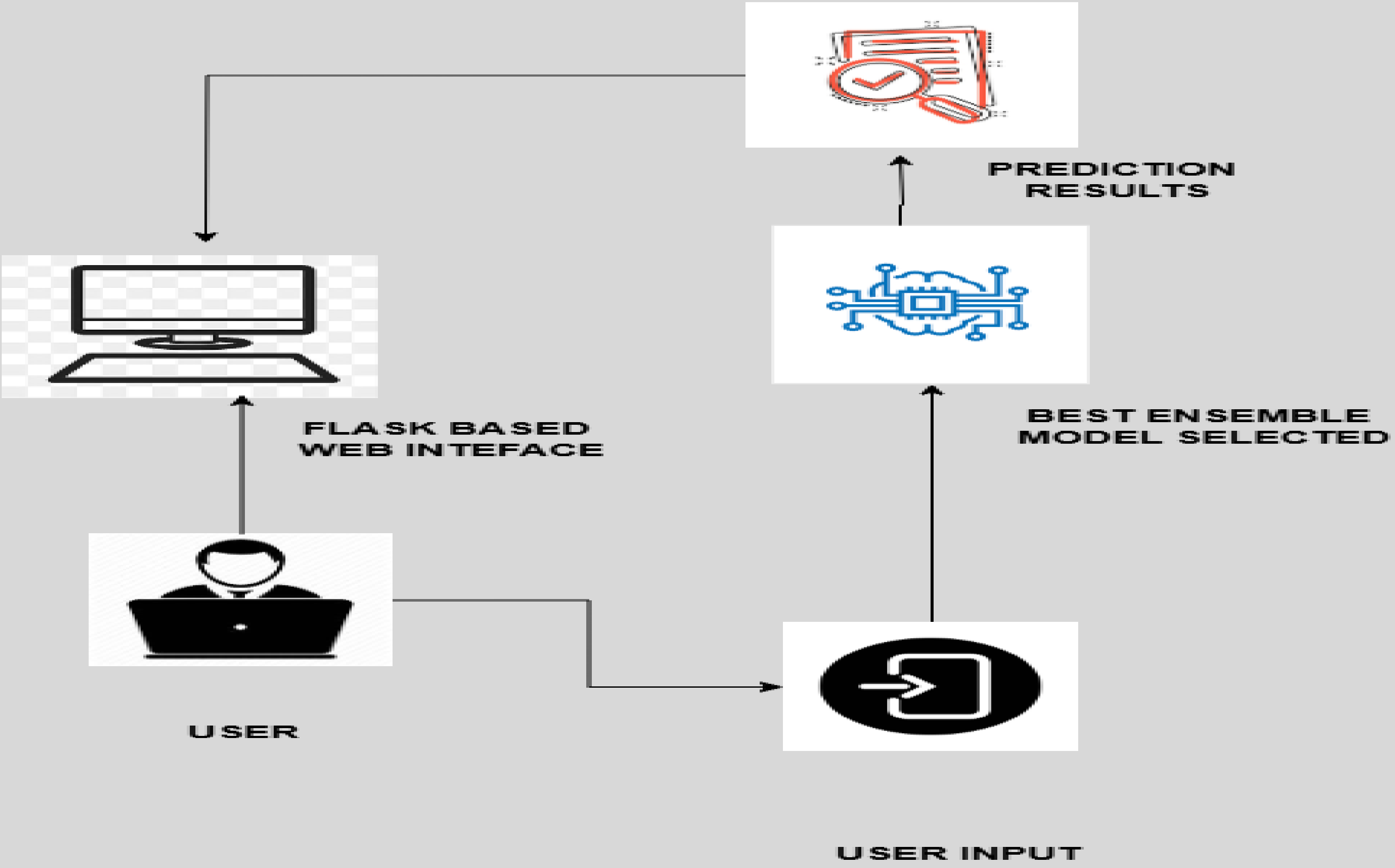
EVALUATION METRICS



COMPARATIVE ANALYSIS

MODELS	EVALUATION METRICS		
	MSE (Mean Square Error)	MAE (Mean Absolute Error)	R2 SCORE
BASE MODELS EVALUATION			
1. Random Forest Regressor	5.2951	0.8859	0.9711
2. Support Vector Regressor	69.2203	6.2239	0.6227
3. XG Boost Regressor	2.9481	0.6484	0.9839
ENSEMBLE MODEL EVALUATION			
1.RNGB (Random Forest Regressor, Gradient Boosting Regressor, Neural Network)	38.6938	3.9048	0.78907
2.RSGB (Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regressor)	1.4832	0.5685	0.9919
3.RXGB (Random Forest Regressor, Gradient Boosting Regressor, XG Boost Regressor)	3.1126	0.6646	0.9830
4.RSXG (Random Forest Regressor, XG Boost Regressor, Support Vector Regressor)	2.9550	0.6486	0.9839

DEPLOYMENT PHASE



PREDICTION RESULTS

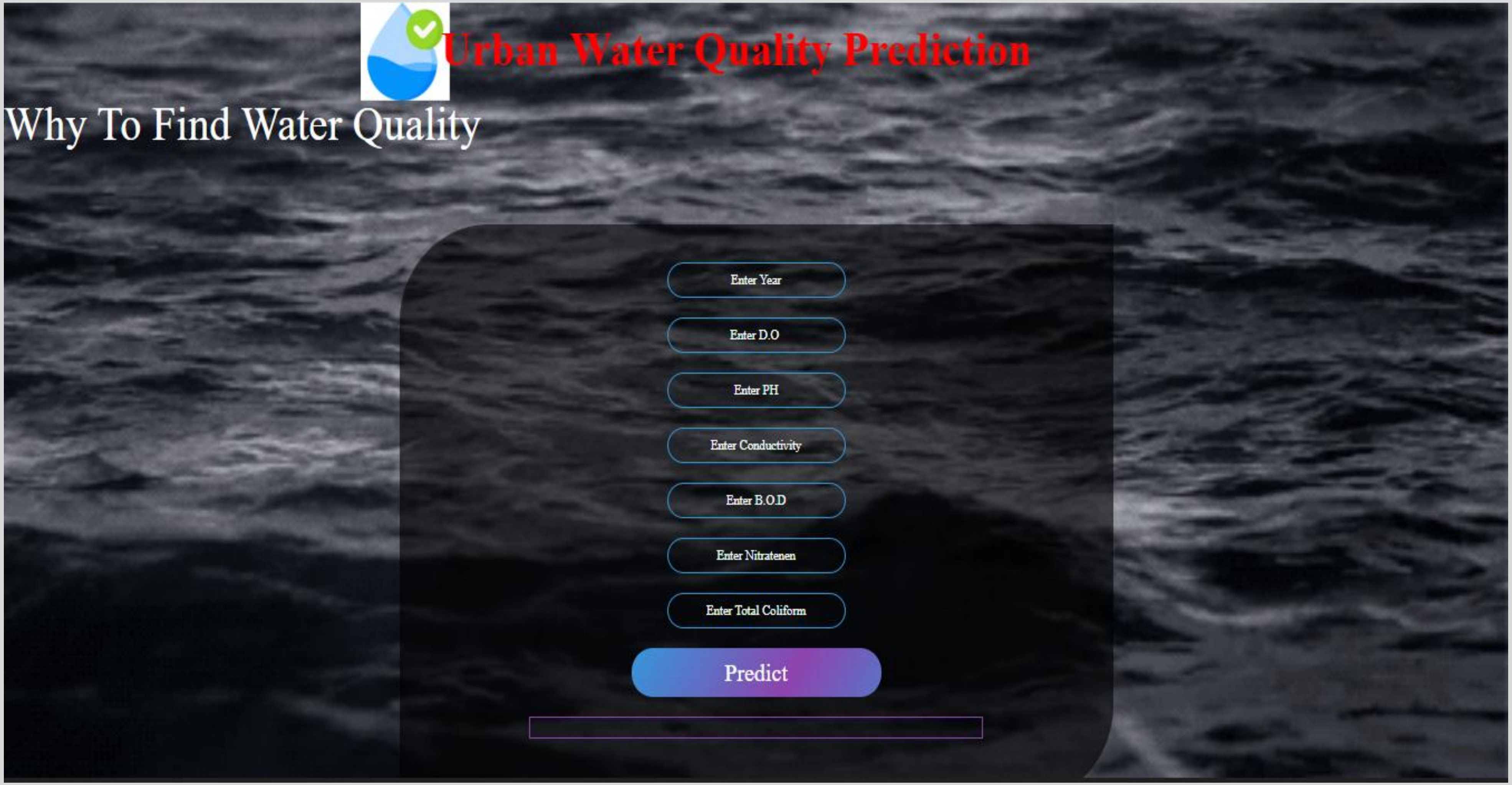
BEST ENSEMBLE MODEL SELECTED

FLASK BASED WEB INTEFACE

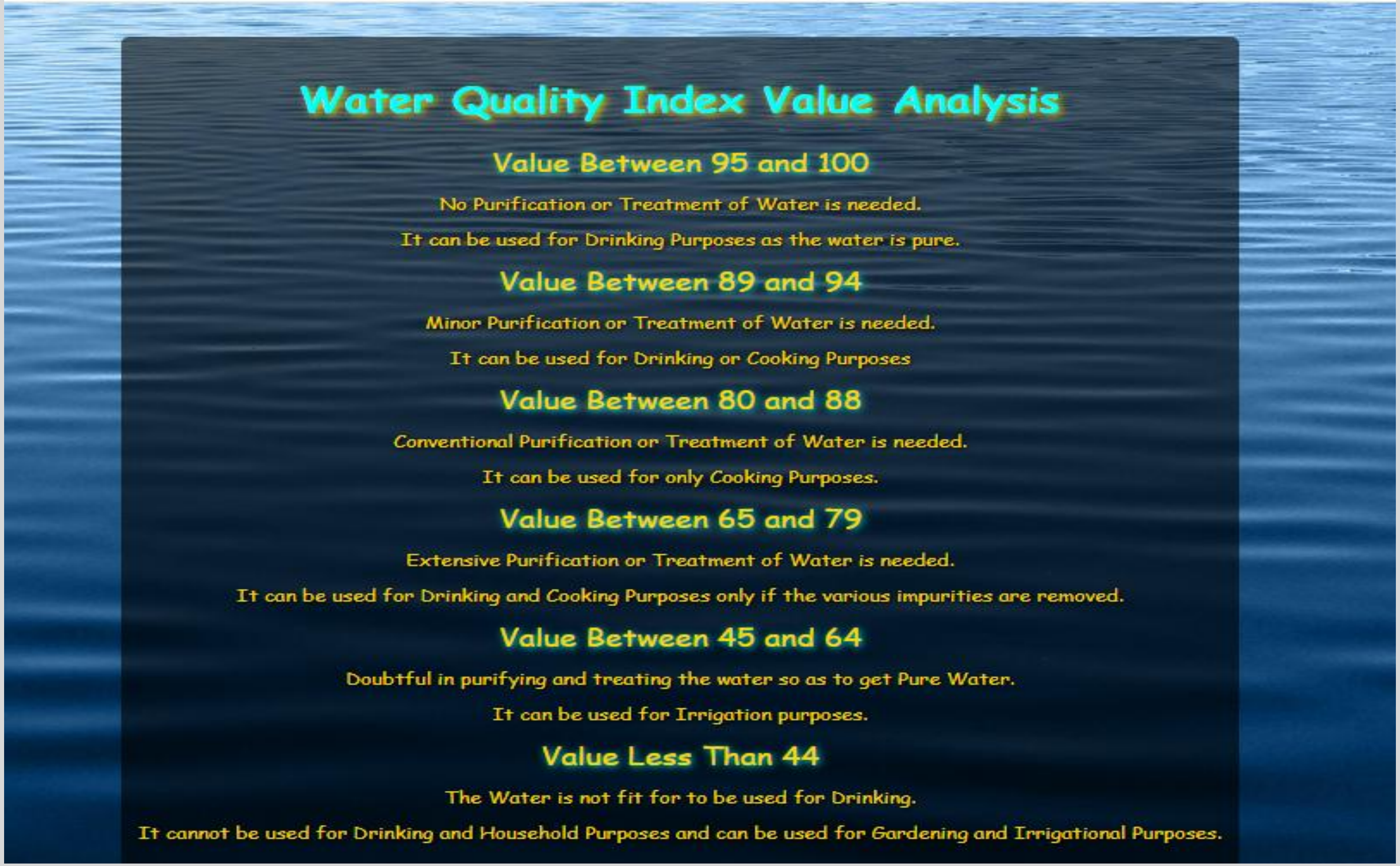
USER

USER INPUT

OUTPUT SCREENSHOTS



OUTPUT SCREENSHOTS



Water Quality Index Value Analysis

Value Between 95 and 100

No Purification or Treatment of Water is needed.

It can be used for Drinking Purposes as the water is pure.

Value Between 89 and 94

Minor Purification or Treatment of Water is needed.

It can be used for Drinking or Cooking Purposes

Value Between 80 and 88

Conventional Purification or Treatment of Water is needed.

It can be used for only Cooking Purposes.

Value Between 65 and 79

Extensive Purification or Treatment of Water is needed.

It can be used for Drinking and Cooking Purposes only if the various impurities are removed.

Value Between 45 and 64

Doubtful in purifying and treating the water so as to get Pure Water.

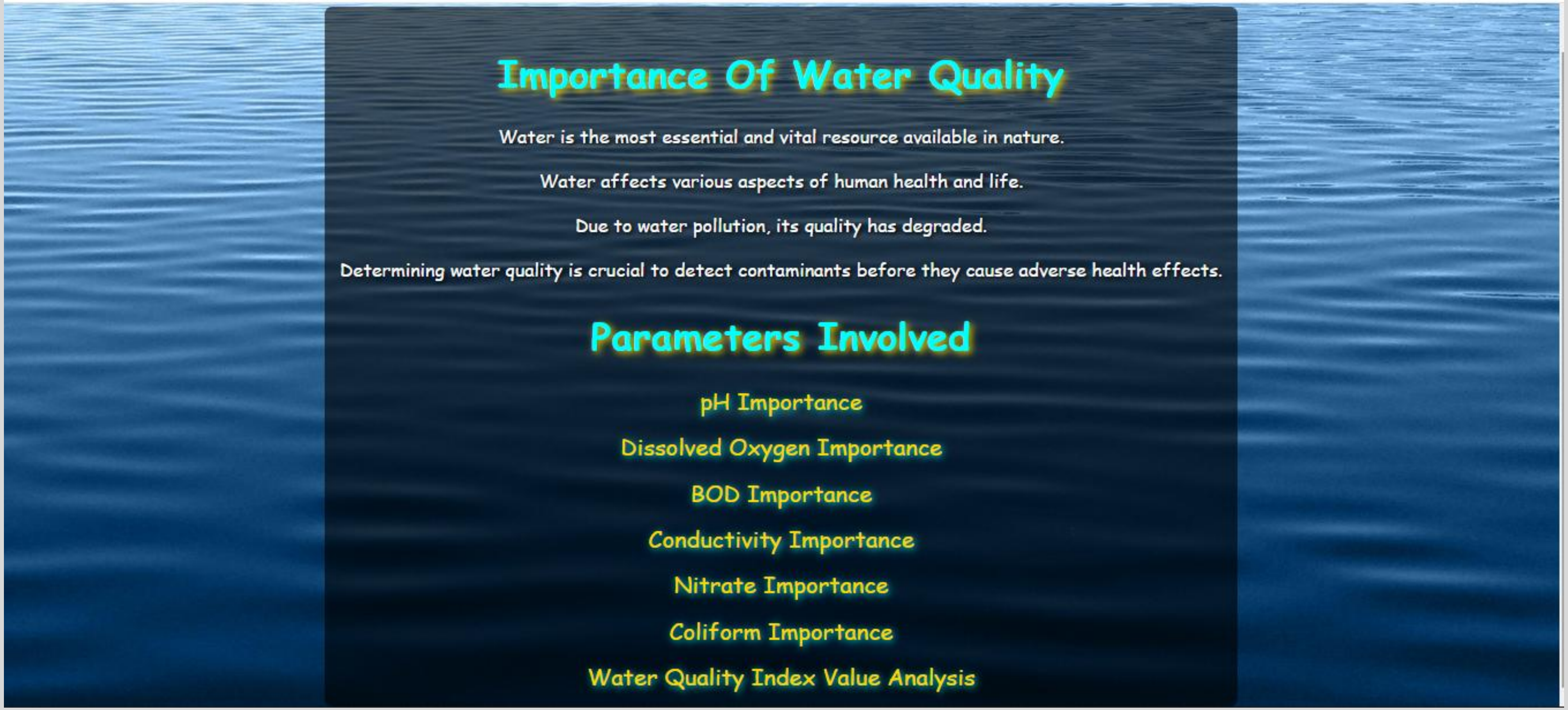
It can be used for Irrigation purposes.

Value Less Than 44

The Water is not fit for to be used for Drinking.

It cannot be used for Drinking and Household Purposes and can be used for Gardening and Irrigational Purposes.

OUTPUT SCREENSHOTS



Importance Of Water Quality

Water is the most essential and vital resource available in nature.

Water affects various aspects of human health and life.

Due to water pollution, its quality has degraded.

Determining water quality is crucial to detect contaminants before they cause adverse health effects.

Parameters Involved

pH Importance

Dissolved Oxygen Importance

BOD Importance

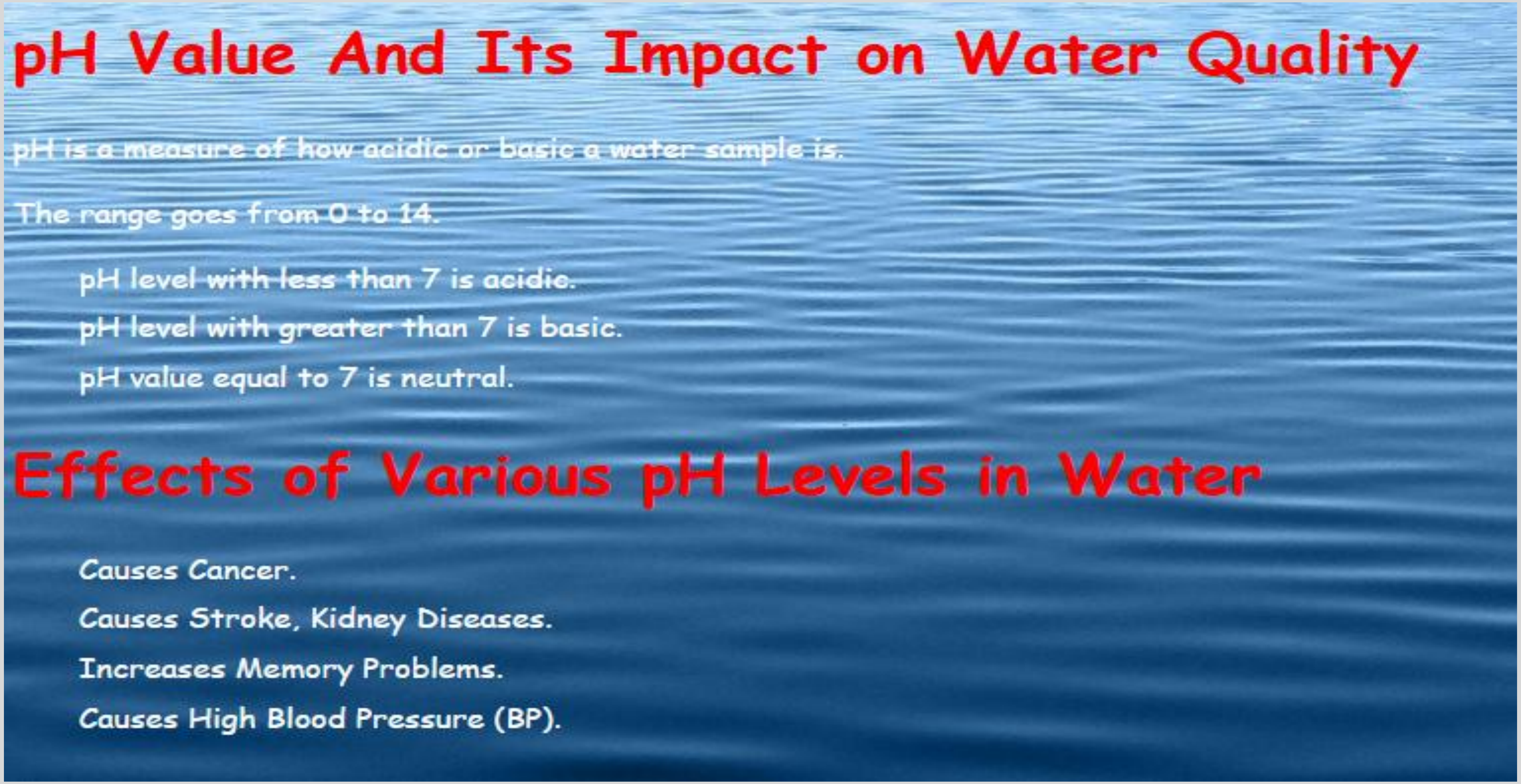
Conductivity Importance

Nitrate Importance

Coliform Importance

Water Quality Index Value Analysis

OUTPUT SCREENSHOTS



pH Value And Its Impact on Water Quality

pH is a measure of how acidic or basic a water sample is.

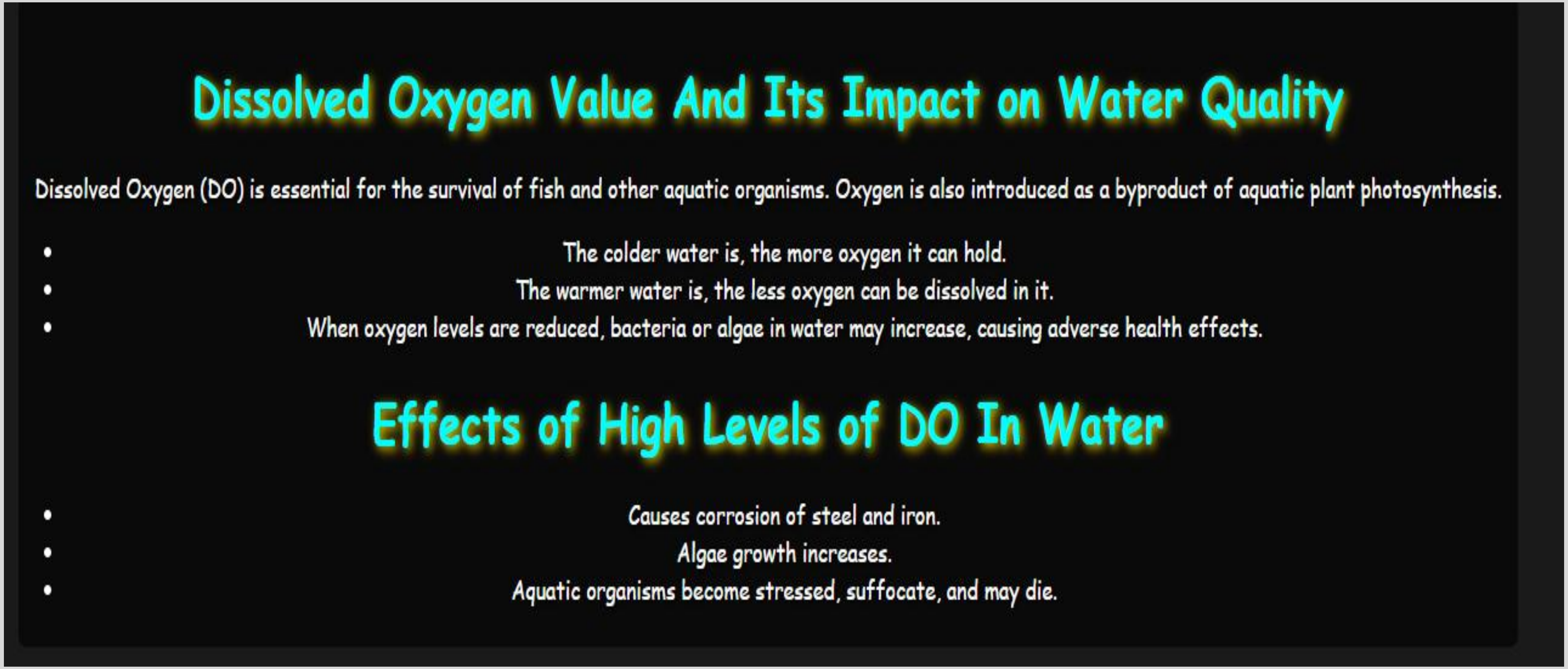
The range goes from 0 to 14.

- pH level with less than 7 is acidic.
- pH level with greater than 7 is basic.
- pH value equal to 7 is neutral.

Effects of Various pH Levels in Water

- Causes Cancer.
- Causes Stroke, Kidney Diseases.
- Increases Memory Problems.
- Causes High Blood Pressure (BP).

OUTPUT SCREENSHOTS



Dissolved Oxygen Value And Its Impact on Water Quality

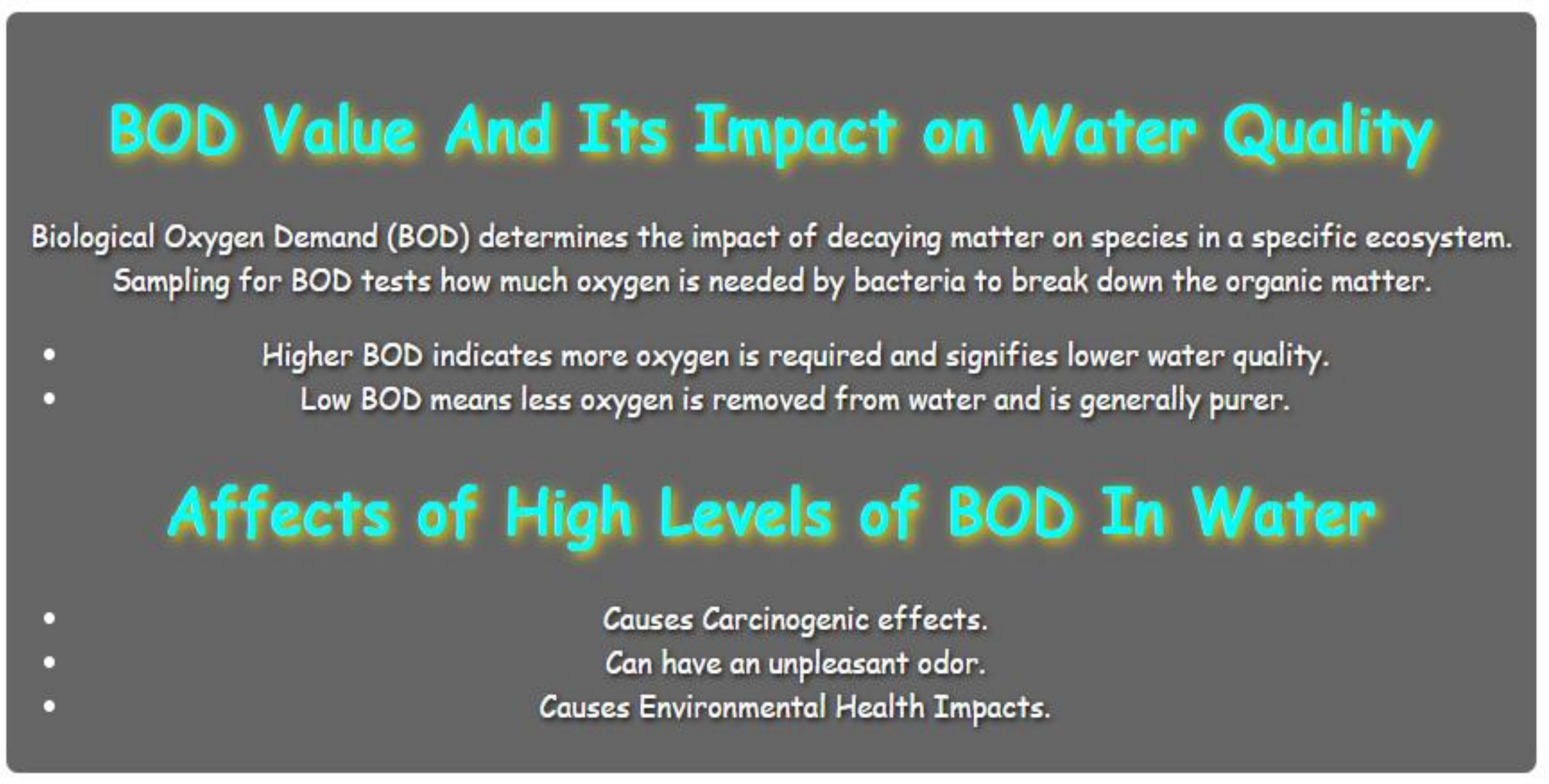
Dissolved Oxygen (DO) is essential for the survival of fish and other aquatic organisms. Oxygen is also introduced as a byproduct of aquatic plant photosynthesis.

- The colder water is, the more oxygen it can hold.
- The warmer water is, the less oxygen can be dissolved in it.
- When oxygen levels are reduced, bacteria or algae in water may increase, causing adverse health effects.

Effects of High Levels of DO In Water

- Causes corrosion of steel and iron.
- Algae growth increases.
- Aquatic organisms become stressed, suffocate, and may die.

OUTPUT SCREENSHOTS



BOD Value And Its Impact on Water Quality

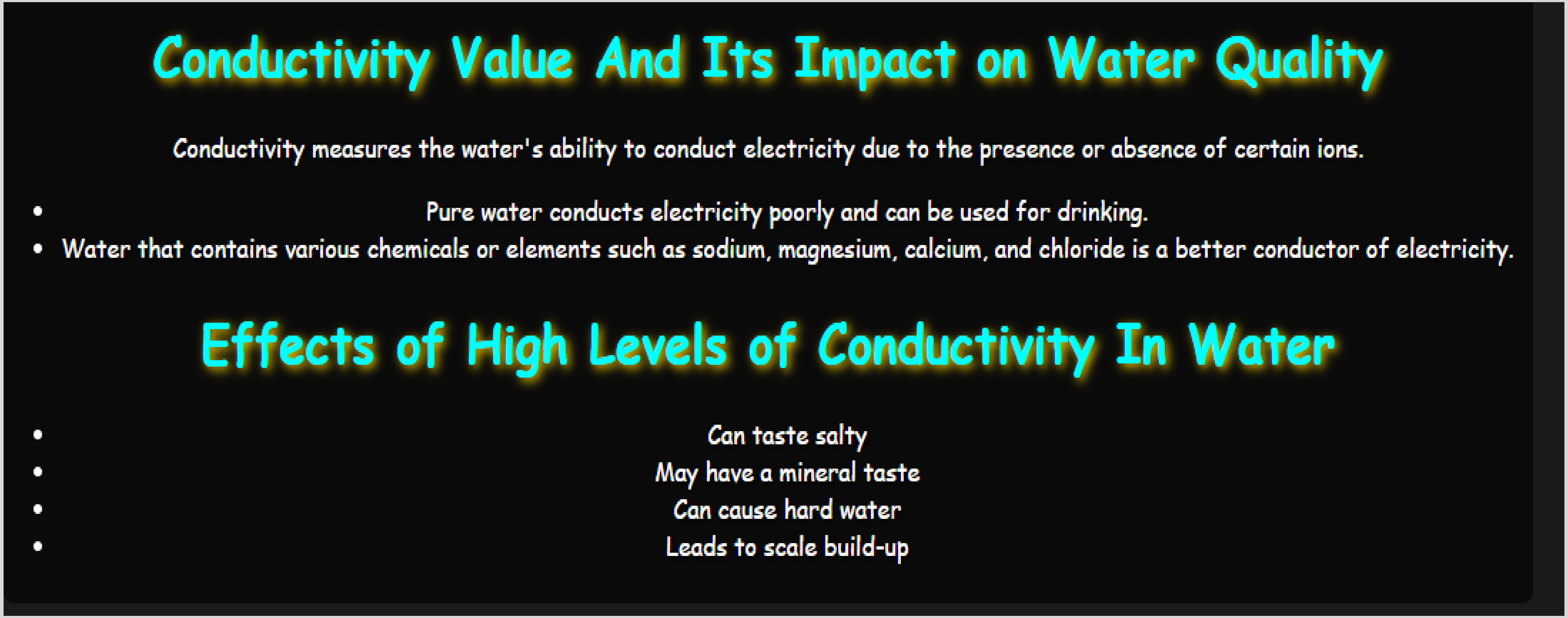
Biological Oxygen Demand (BOD) determines the impact of decaying matter on species in a specific ecosystem. Sampling for BOD tests how much oxygen is needed by bacteria to break down the organic matter.

- Higher BOD indicates more oxygen is required and signifies lower water quality.
- Low BOD means less oxygen is removed from water and is generally purer.

Affects of High Levels of BOD In Water

- Causes Carcinogenic effects.
- Can have an unpleasant odor.
- Causes Environmental Health Impacts.

OUTPUT SCREENSHOTS



Conductivity Value And Its Impact on Water Quality

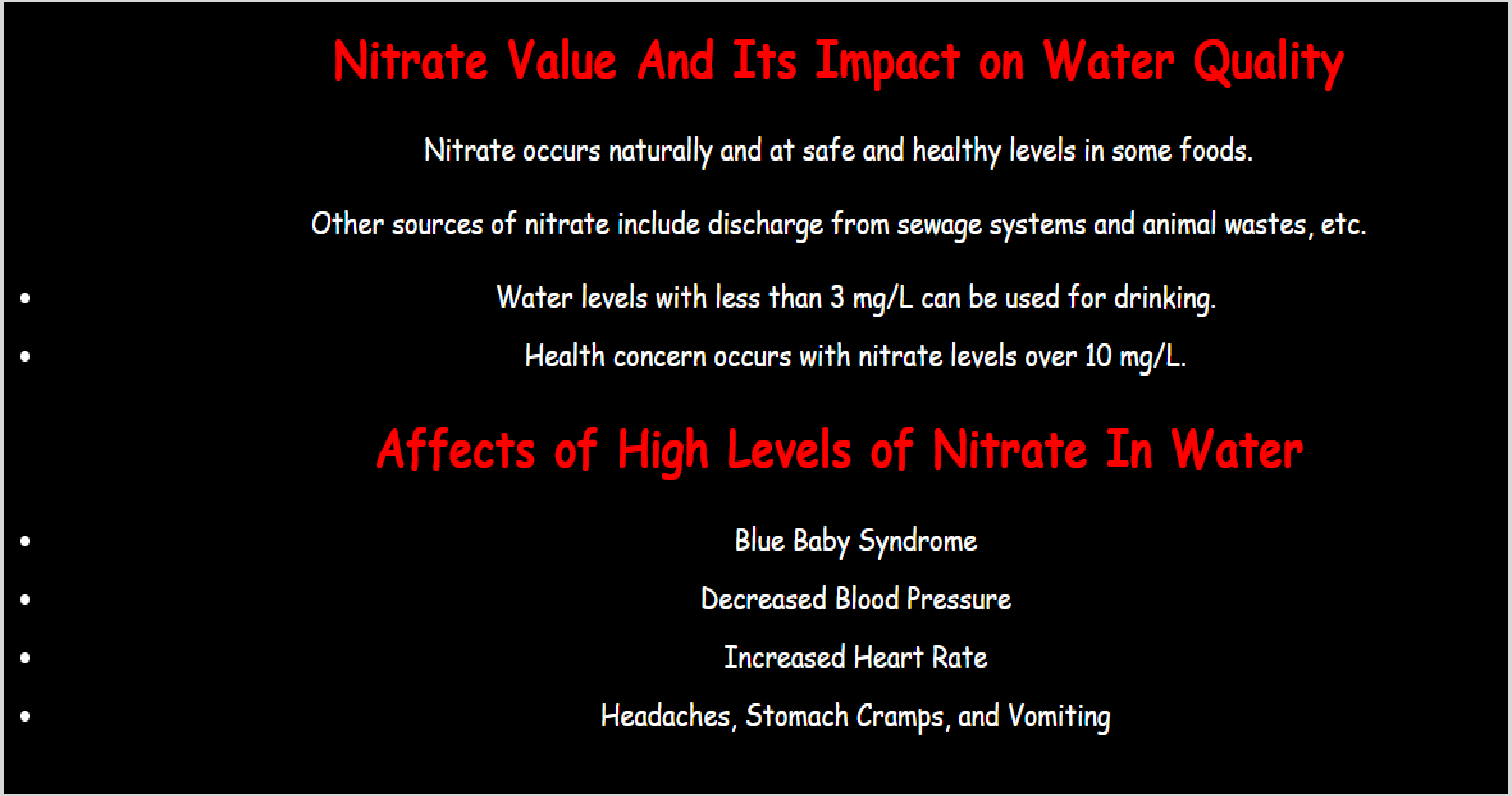
Conductivity measures the water's ability to conduct electricity due to the presence or absence of certain ions.

- Pure water conducts electricity poorly and can be used for drinking.
- Water that contains various chemicals or elements such as sodium, magnesium, calcium, and chloride is a better conductor of electricity.

Effects of High Levels of Conductivity In Water

- Can taste salty
- May have a mineral taste
- Can cause hard water
- Leads to scale build-up

OUTPUT SCREENSHOTS



Nitrate Value And Its Impact on Water Quality

Nitrate occurs naturally and at safe and healthy levels in some foods.

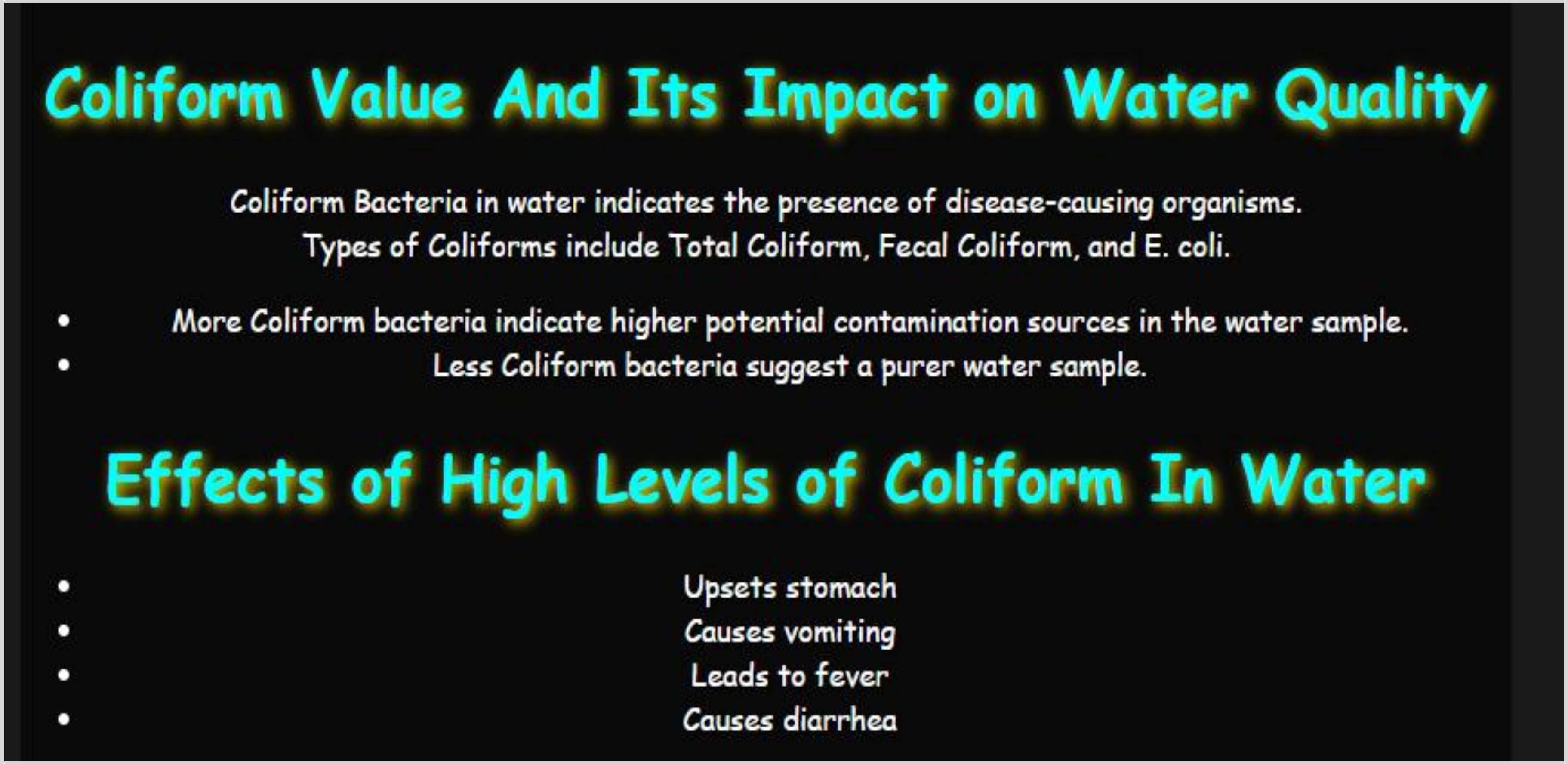
Other sources of nitrate include discharge from sewage systems and animal wastes, etc.

- Water levels with less than 3 mg/L can be used for drinking.
- Health concern occurs with nitrate levels over 10 mg/L.

Affects of High Levels of Nitrate In Water

- Blue Baby Syndrome
- Decreased Blood Pressure
- Increased Heart Rate
- Headaches, Stomach Cramps, and Vomiting

OUTPUT SCREENSHOTS



Coliform Value And Its Impact on Water Quality

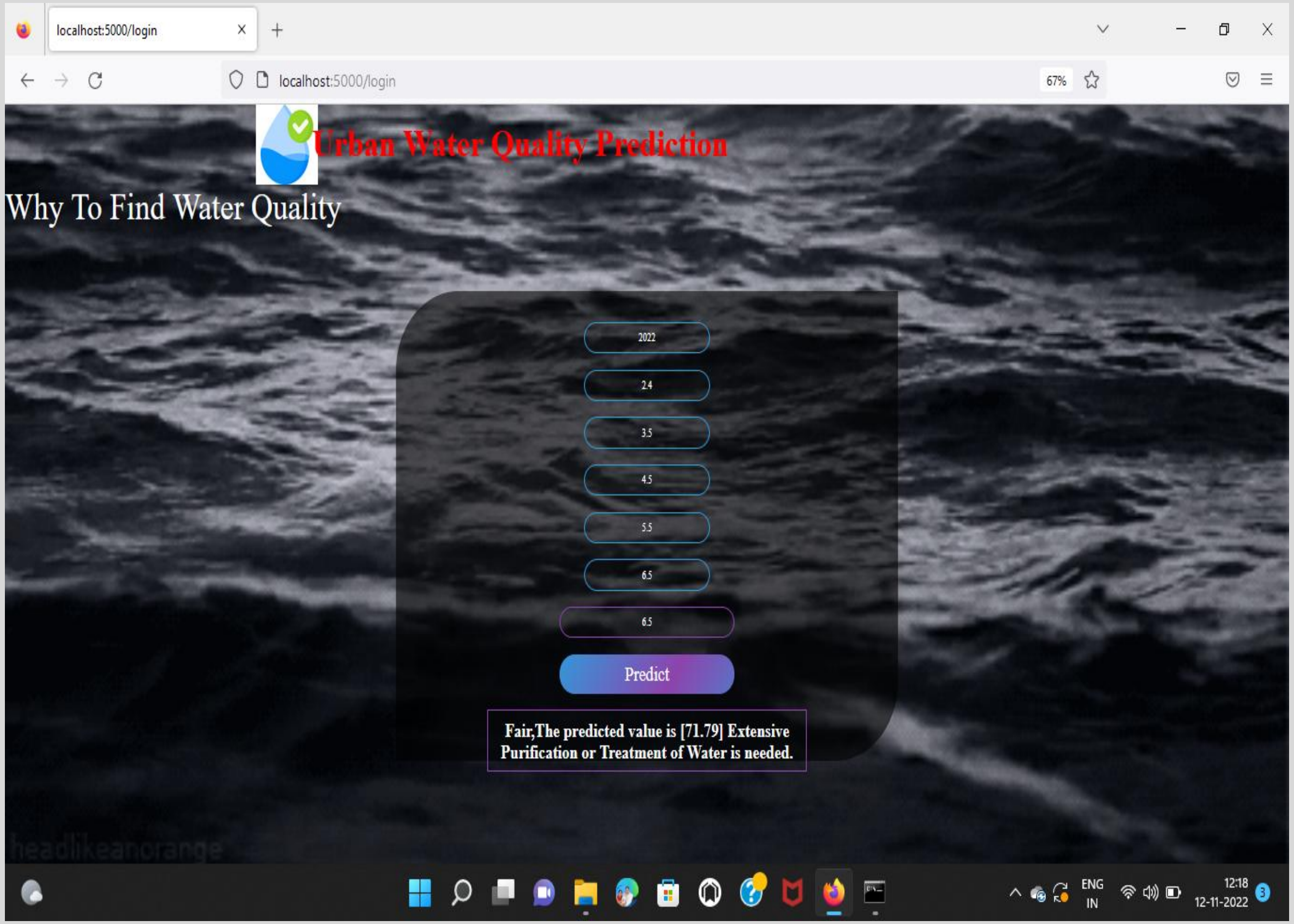
Coliform Bacteria in water indicates the presence of disease-causing organisms.
Types of Coliforms include Total Coliform, Fecal Coliform, and E. coli.

- More Coliform bacteria indicate higher potential contamination sources in the water sample.
- Less Coliform bacteria suggest a purer water sample.

Effects of High Levels of Coliform In Water

- Upsets stomach
- Causes vomiting
- Leads to fever
- Causes diarrhea

OUTPUT SCREENSHOTS



2022

2.4

3.5

4.5

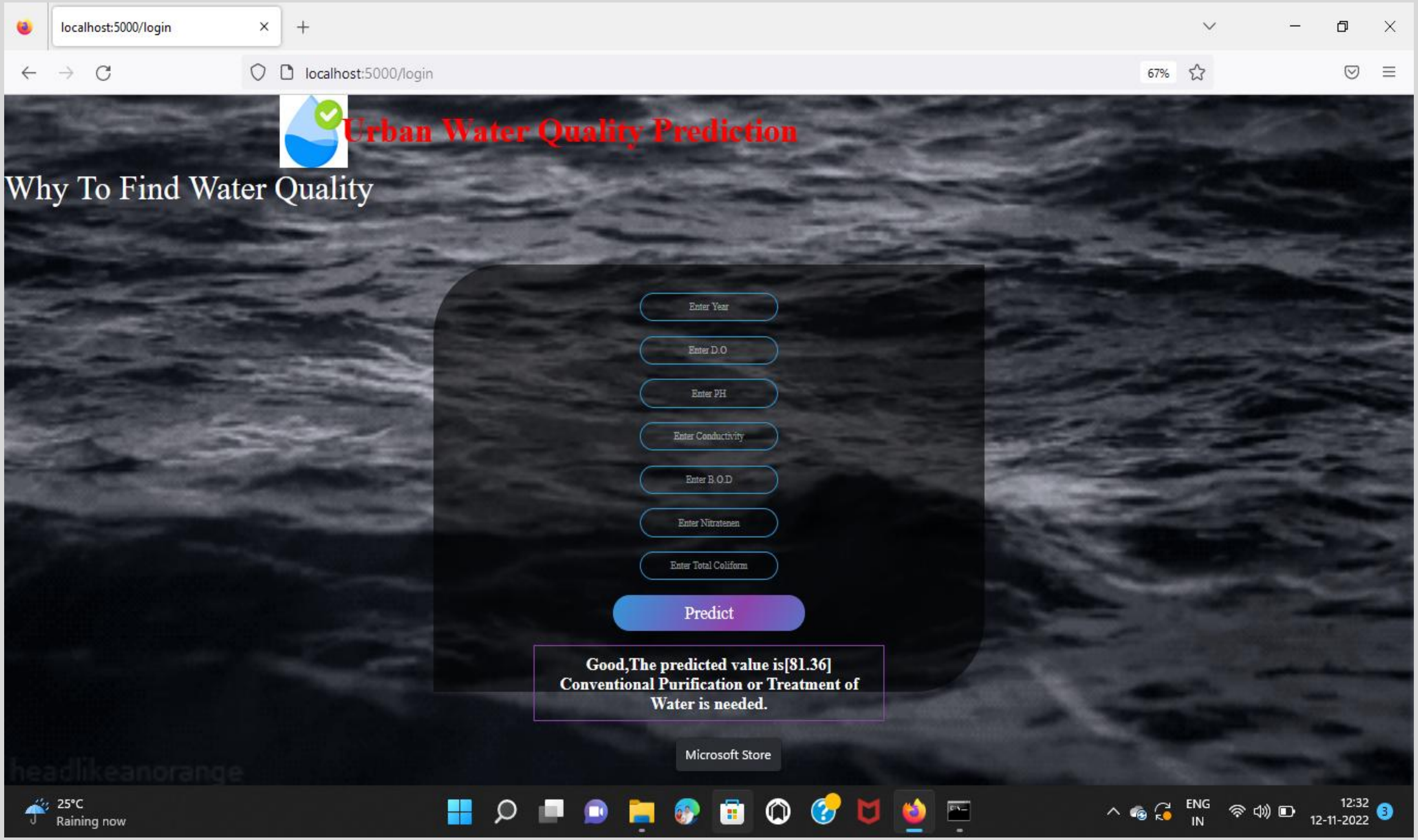
5.5

6.5

6.5

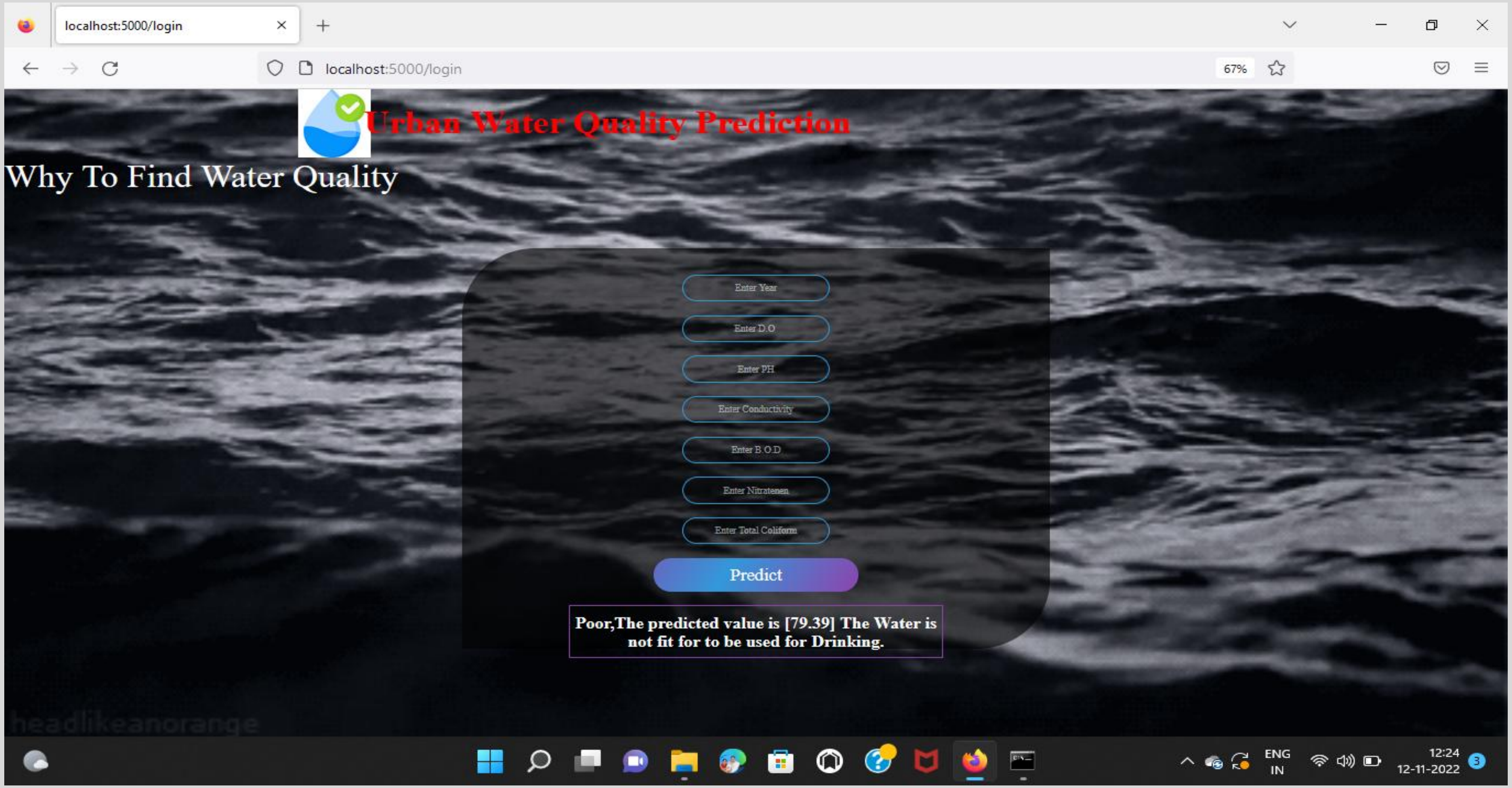
Fair,The predicted value is [71.79] Extensive Purification or Treatment of Water is needed.

OUTPUT SCREENSHOTS



Good,The predicted value is[81.36] Conventional Purification or Treatment of Water is needed.

OUTPUT SCREENSHOTS



Poor,The predicted value is [79.39] The Water is not fit for to be used for Drinking.

CONCLUSION

- This work demonstrates an effective approach for water quality prediction by combining machine learning with advanced feature extraction and ensemble modelling.

- By leveraging several ensemble model combinations and its comparative analysis, along with domain informed feature engineering (RFE) and SHAP analysis, the work achieves high interpretability and elevated the analysis process involved in water quality.

- This solution provides valuable insights into water quality trends, supporting proactive management and safeguarding of water resources for public health and environmental sustainability.

THANK YOU