International Journal of Research Publication and Reviews

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

Prediction of Drinking Water Potability using Machine Learning

Embari Sahith [a], Bethi Sravani [b], Telugu Hariprasad [c], Kura Sai Kiran [d], B.Siris Royal [e]*

[a,b,c,d] Student, Department of AIML. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

[e] Professor, Department of IT. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

ABSTRACT

Water is a fundamental resource for sustaining human life, and maintaining its cleanliness is crucial for individual health. Contaminated drinking water poses substantial health risks, including diseases such as diarrhoea, cholera, and various other waterborne infections. Thus, ensuring the availability of safe and clean water is crucial for promoting public health. Recent data indicate that over 3,575,000 individuals die each year due to waterborne illnesses. Thus, accurate prediction of water potability could substantially reduce the incidence of such illnesses. Machine learning algorithms have emerged as powerful tools for properly forecasting water quality, enabling timely and precise monitoring of water resources. This project investigates different techniques for predicting water potability based on the physicochemical properties of water samples obtained from the Drinking Water dataset on Kaggle. This dataset has nine distinct parameters: pH, hardness, solids, chloramines, sulphates, trihalomethanes, organic carbon, conductivity, and turbidity. We employ various methods, including Random Forest, Logistic Regression, Support Vector Classification, and K-Nearest Neighbours, to evaluate the potability of drinking water. The Random Forest method exhibits superior performance compared to traditional machine learning models, with an accuracy of 81%. Additionally, the Support Vector Classifier (SVC) achieves a commendable accuracy of 68%. This research holds significant promise in providing reliable water quality data to scholars, water management authorities, and policymakers, hence enhancing the effectiveness of water potability assessment.

Keywords - Water potability prediction, Random Forest, SVC, KNN, Logistic Regression

I. INTRODUCTION

Water, an essential resource for all life on Earth, holds considerable significance in the realms of economy, environment, and human health. The provision of safe and uncontaminated drinking water is essential for preserving health and averting waterborne infections. Contaminated water can include dangerous bacteria, viruses, parasites, and chemicals, resulting in several diseases and infections, including diarrhoea, dysentery, typhoid, polio, cholera, and hepatitis A. The WHO estimates that approximately 485,000 persons perish annually from diarrhea-related complications attributable to contaminated drinking water. Furthermore, polluted water sources might precipitate chronic health conditions, such as cancer and developmental abnormalities. In India, alarming statistics indicate that more than 37.7 million individuals, including children, are afflicted by waterborne diseases each year. Disturbingly, domestic and industrial pollution has contaminated over 70% of available water, leaving approximately 80% of the rural population and 20% of the urban population without access to safe drinking water. Global society has significant challenges about water scarcity and deteriorating water quality, adversely affecting millions worldwide. The World Health Organization's 2018 report indicates that about 2 billion individuals are subjected to water contaminated with faecal matter. Consequently, to achieve sustainable development, promote a healthy lifestyle, and eradicate poverty, ensuring everyone access to clean and safe water is essential.

In areas lacking adequate water treatment facilities, such as developing nations and rural locales, the ability to predict water potability is crucial. Traditional methods for evaluating water quality, marked by costly and labor-intensive laboratory and statistical investigations, have shown to be inefficient. Thus, there is an immediate necessity for a more effective and cost-efficient alternative. This research aims to propose and assess the viability of a machine learning approach for the real-time prediction of water potability. In recent years, machine learning algorithms have made significant advancements in predicting water quality, enabling more precise and efficient monitoring. Various classification techniques, such as Support Vector Classification (SVC), Random Forest (RF), K-Nearest Neighbours (KNN), and Logistic Regression, can be employed to assess water potability.

The primary focus is on predicting the potability of drinking water based on its physicochemical characteristics. The study is to develop a model that provides accurate and timely information on drinking water quality, enabling policymakers and water resource managers to implement preventive measures and ensure public access to clean drinking water. The study aims to assess the efficacy and precision of several algorithms used in forecasting water quality.

II. LITERATURE SURVEY

Phorah et al. (2023)[1] emphasised the significance of data preparation techniques in enhancing the efficacy of machine learning models for forecasting water potability. Their research focused on several strategies such as data purification, feature selection, and dimensionality reduction to improve model accuracy and effectiveness. Darmawan et al. (2024)[2] utilised an Artificial Neural Network (ANN) methodology to classify the potability of drinking water. Their research investigated how deep learning techniques could enhance water quality assessment and provide more accurate predicted insights. Patel et al. (2022) advocated for the application of Random Forest and Gradient Boosting algorithms to assess water potability, specifically for Indian rivers. Their research highlighted the effectiveness of ensemble learning techniques in handling complex datasets related to water quality. Aldhyani et al. (2020) employed Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) techniques for forecasting water quality. Their research demonstrated the efficacy of deep learning models in analysing temporal patterns in water quality data. Al Duhayyim et al. (2022) introduced a fuzzy deep neural network optimised using an atom search algorithm for the prediction of water quality. Their research improved the robustness of water quality classification models. Rustam et al. (2022) developed Nonlinear AutoRegressive Neural Network (NARNET) and Long Short-Term Memory (LSTM) models for the prediction of water quality. Their research provided insights into time-series forecasting techniques for assessing water potability. Ozsezer & Mermer (2023) performed a comparative examination of various machine learning algorithms, concluding that XGBoost and Random Forest are the most effective for predicting drinking water quality. Their findings underscored the importance of algorithm selection in improving predictive accuracy. Alnaqeb et al. (2022)[8] shown the effectiveness of the LightGBM algorithm in assessing water potability. Their research illustrated the effectiveness of gradient boosting techniques in handling imbalanced water quality datasets.

Study	**Key Contribution**	**Year**
Phorah et al.	Emphasized advanced data preprocessing techniques to enhance machine learning model performance in water potability prediction.	2023
Darmawan et al.	Applied an Artificial Neural Network algorithm for classifying water potability.	2024
Patel et al.	Proposed the use of Random Forest and Gradient Boost techniques for predicting water potability in Indian rivers.	2022
Aldhyani et al.	Employed LSTM and SVM techniques for water quality prediction.	2020
Al Duhayyim et al.	Proposed a fuzzy deep neural network with atom search optimization for water quality prediction.	2022
Rustam et al.	Developed NARNET and LSTM models for water quality prediction.	2022
Ozsezer & Mermer	Compared various machine learning algorithms, identifying XGBoost and Random Forest as effective methods for predicting drinking water quality.	2023
Alnaqeb et al.	Demonstrated the effectiveness of LightGBM for assessing water potability.	2022

III. METHODOLOGY

The methodology for this research focuses on leveraging machine learning techniques, statistical analysis, and data-driven approaches to predict the drinking water potability. The process involves several key steps, including data collection, preprocessing, feature engineering, model selection, and evaluation.

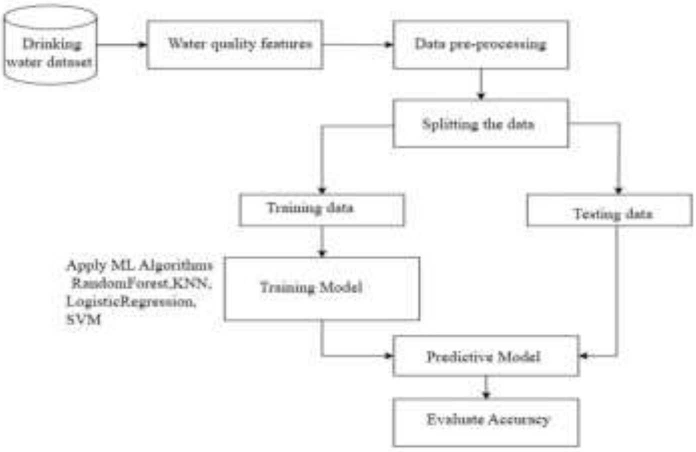
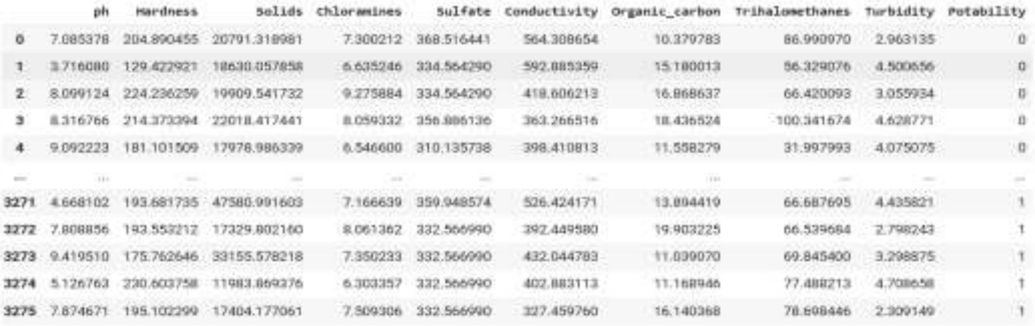


Fig 1.**System Architecture**

3.1. Data Collection

The first step involves gathering a dataset containing physicochemical properties of water samples.The proces starts with a water quality dataset, which contains various features related to water properties, such as pH, Hardness,Solids,Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity, and Potability.Potability (0 or 1) is the target variable,where1 means potable (safe to drink) and 0 means non-potable.



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	7.085378	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	334.564290	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	334.564290	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	332.566990	392.449580	19.903225	66.539684	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	332.566990	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	332.566990	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	332.566990	327.459760	16.140368	78.698446	2.309149	1

Fig 2. **Sample Dataset**

3.2. Data Preprocessing

The second step of this approach is to clean the dataset, which means fixing bad data such as empty cells, data in the wrong format, wrong data and duplicates. In our dataset, the number of instances for potable samples (represented by 1) is 1200, and the number of instances for non-potable samples (represented by 0) is 811. In this case, the majority of samples belong to one class (1, potable water), which means that the raw dataset is imbalanced and the accuracy and other scores can be misleading because the model will predict everything as the majority class and will still achieve high accuracy even if it is not accurate. To solve this issue, undersampling is used by reducing the number of potable samples to match the number of non-potable samples. There fore, a balanced dataset was created by alternating between the two potability classes and maintaining the same number of potabilities 0 and 1.

After data cleaning, the dataset is methodically partitioned into three distinct segments (Table 1).

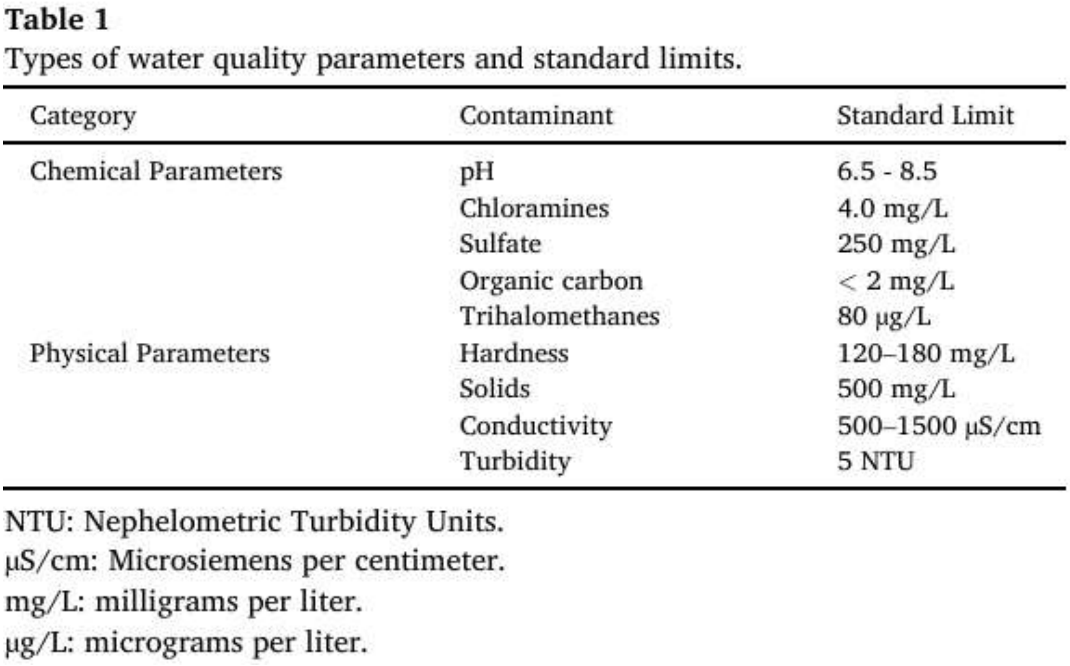


Table 1

Types of water quality parameters and standard limits.

Category	Contaminant	Standard Limit
Chemical Parameters	pH	6.5 - 8.5
	Chloramines	4.0 mg/L
	Sulfate	250 mg/L
	Organic carbon	< 2 mg/L
	Trihalomethanes	80 μg/L
Physical Parameters	Hardness	120–180 mg/L
	Solids	500 mg/L
	Conductivity	500–1500 μS/cm
	Turbidity	5 NTU

NTU: Nephelometric Turbidity Units.
μS/cm: Microsiemens per centimeter.
mg/L: milligrams per liter.
μg/L: micrograms per liter.

***Table 1*.water quality parameters and standards**

3.3. Feature Engineering

Feature engineering enhances machine learning models for drinking water potability prediction by extracting meaningful insights from water quality parameters. Key features include pH categorization (acidic, neutral, alkaline) for acidity assessment, analysis of hardness and total dissolved solids (TDS) for mineral content impact, and chemical concentrations (sulfate, chloramines, trihalomethanes) to detect contamination and treatment effectiveness.

3.4. Model Training

To ensure accurate drinking water potability predictions, multiple machine learning models were trained and evaluated. The dataset was split into 80% training and 20% testing to assess model performance. The following models were implemented:

3.4.1 Logistic Regression

Logistic Regression is a classification algorithm used to predict water potability. It models the probability that a given water sample is potable using the sigmoid function:

$P(Y=1/X) =1/1+e-(wX+b)$

where $P(Y=1/X)$ is the probability of the water being potable, and w and b are model coefficients.

3.4.2 Support Vector Machine (SVM)

SVM is used to classify water as potable or non-potable by finding an optimal hyperplane that maximizes the margin between classes. The decision function is given by:

$f(X)=wX+b$

where w is the weight vector, X is the feature set, and b is the bias term.

3.4.3 K Nearest Neighbors (KNN)

KNN classifies water samples based on their similarity to the nearest neighbors. The Euclidean distance formula determines the closest points:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Where d represents the distance between two points.

3.4.4 Random Forest

Random Forest is an ensemble model using multiple decision trees. It classifies water samples based on majority voting among trees:

$$f(X) = \frac{1}{N}\sum_{i=1}^{N} h_i(X)$$

Where $h_i(X)$ is the prediction from each decision tree and N is the number of trees.

3.5 Model Evaluation

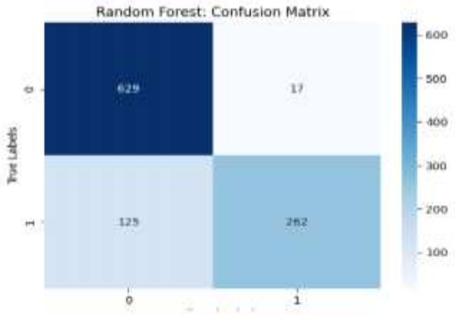
The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Metric	**Formula**
Precision (P)	(TP) / (TP + FP)
Recall (R)	(TP) / (TP + FN)
Accuracy	(TP+TN)/(TP+TN+FN+FP)
F1-score	2*(Recall*Precision)/(Recall+Precision)

Table 2 **.The performance metrics used for classification and regression**

V. RESULT ANALYSIS AND DISCUSSION

The dataset consisted of 3,276 water samples with 9 physicochemical attributes, providing a diverse set of features for training machine learning models. The data was split into an 80:20 training-to-testing ratio, ensuring a balanced evaluation of model performance. The Random Forest model outperformed all other algorithms, demonstrating the highest accuracy in predicting water potability, followed closely by SVM and KNN. Logistic Regression had the lowest accuracy, highlighting its limitations in handling complex dependencies among water quality parameters.



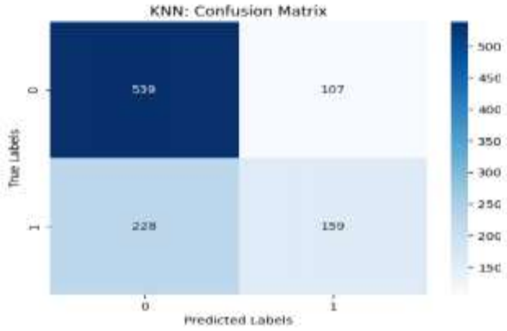
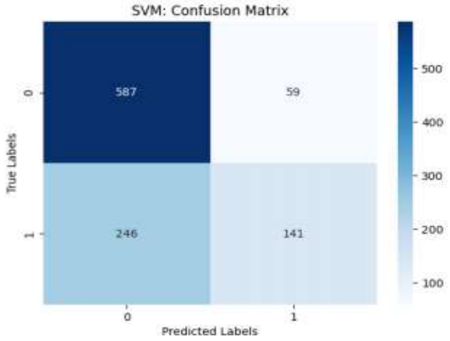


Fig 3.**confusion matrix for Random Forest**

Fig 4.**confusion matrix for KNN**



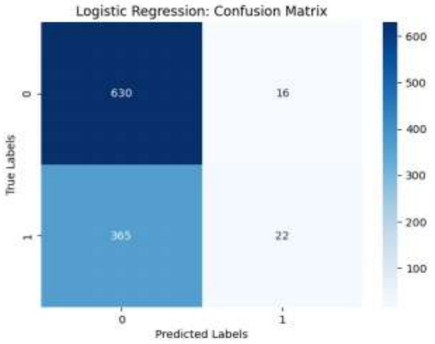


Fig 5. **confusion matrix for SVM**

Fig 6.**confusion matrix for Logistic Regression**

The confusion matrices for the implemented models provide insights into their predictive performance for drinking water potability. Support Vector Machine (SVM) demonstrates moderate accuracy, correctly identifying most non-potable water instances but struggling with a high number of false negatives, indicating difficulty in detecting potable water. Random Forest outperforms other models, exhibiting a low number of false positives and false negatives, making it a reliable choice for water potability prediction. K-Nearest Neighbors (KNN), however, shows weaker performance, with a significant number of misclassifications, particularly false positives and false negatives, suggesting sensitivity to variations in data. Logistic Regression performs the worst, with an extremely high number of false negatives, indicating that it fails to identify potable water correctly. Among these models, Random Forest proves to be the most effective, whereas Logistic Regression struggles to provide accurate predictions. For the proposed research, a dataset consisting of 3276 samples was employed. Each sample underwent analysis to determine nine specific water quality parameters: pH, Organic_carbon, Chloramines, Turbidity, Trihalomethanes, Sulphate, Hardness, Conductivity and Solids. A summary of these parameters is provided in Table 3. To facilitate the analysis, the dataset was divided into 80:20 ratio of training and test data.

Potable	**Non potable**
39.01%	63.99%

***Table 3.* Percentage of potable and non-potable water based on the dataset**

This study aimed to evaluate the performance of various algorithms such as logistic regression, KNN, RF, and SVC by employing multiple performance metrics via confusion matrix.

Model Name	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression (LR)	62.8	62.8	100	77.2
Random Forest Classifier	81.09	79.4	87.6	77.5
Support Vector Classifier (SVC)	68.8	62.8	100	77.2
K-Nearest Neighbors (KNN)	63.5	61.6	71.6	66.2

***Table 4.* Proposed algorithms performance analysis**

Algorithm	Accuracy(%)	Key Characteristics
Logistic Regression	62.65	Works well with linearly separable data; interpretable but limited for complex relationships
K-Nearest Neighbors	63.57	Simple, non-parametric; sensitive to the choice of K and distance metric.
Random Forest	81.10	Ensemble learning technique; reduces overfitting; computationally expensive.
Support Vector Classifier (SVC)	68.29	Effective for small datasets; performs well with high-dimensional data

***Table 5.* Comparative Summary of Models**

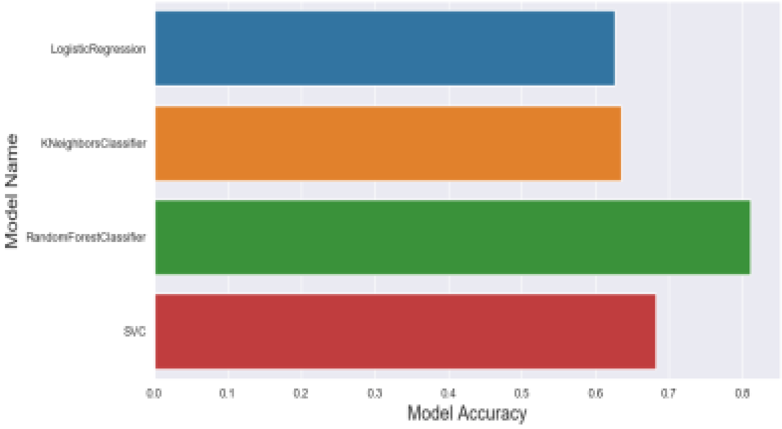


***Fig 7.* Accuracy comparison among the models**

The research presented the accuracy scores of four machine learning algorithms, visualized in the bar graph shown in Fig 7 Based on the graph, it can be inferred that Random Forest achieved the highest accuracy rate, followed by SVM, and KNN. In contrast, the Logistic Regression model demonstrated the lowest accuracy rate among the algorithms.

VI. CONCLUSION

Safeguarding the safety and purity of drinking water is essential for the protection of human health. Accurate prediction of water potability is crucial for achieving this objective. Access to drinkable water is an intrinsic right for all humans, vital for maintaining overall health and preventing waterborne diseases. The increasing global population and escalating pollution levels have raised significant concerns about the quality of water supplies. Employing machine learning techniques can substantially assist in predicting water potability and implementing necessary measures to enhance water quality, hence guaranteeing the supply of safe drinking water for the population. A recent study meticulously analysed various machine learning algorithms and their effectiveness in predicting water potability based on an extensive set of physicochemical parameters. This research demonstrates the effectiveness of all employed algorithms as vital tools for monitoring and managing water quality, with substantial implications for the water sector and public health. It is essential to acknowledge particular limitations of the study. The dataset utilised in the study is rather constrained, consisting of 3,276 observations. Consequently, extrapolating the findings to larger groups may be challenging. The research focused on a limited set of water quality parameters, and it is advised that future investigations include additional relevant factors that may influence water potability.

References

[1]. Darmawan, I., Fatchan, M., & Firmansyah, A. (2024). Classification of Drinking Water Potability With Artificial Neural Network Algorithm. International Journal of Integrated Science and Technology, 2(5), 506–515.

[2]. Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., Althobaiti, Y. S., & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Computational Intelligence and Neuroscience, 2022, 9283293.

[3]. Mohan, B. R., Dileep, M., Bhuria, V., Gadde, S. S., Kumarasamy, M., & Prasad, A. N. (2023). Potable Water Identification with Machine Learning: An Exploration of Water Quality Parameters. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 178–185.

[4]. Ainapure, B., Baheti, N., Buch, J., Appasani, B., Jha, A. V., & Srinivasulu, A. (2023). Drinking water potability prediction using machine learning approaches: a case study of Indian rivers. Water Practice and Technology, 18(12), 3004–3020.

[5]. Phorah, K., Sumbwanyambe, M., & Sibiya, M. (2024). Systematic Literature Review on Data Preprocessing for Improved Water Potability Prediction: A Study of Data Cleaning, Feature Engineering, and Dimensionality Reduction Techniques. Nano Progress, 20(S11)

[6]. Patel, S., Patel, D., & Patel, R. (2023). Water Potability Prediction Using Machine Learning. ResearchGate.

[7]. Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M., "Water quality prediction using artificial intelligence algorithms", Applied Bionics and Biomechanics, 2020

[8]. Addisie, M.B., "Evaluating Drinking Water Quality Using Water Quality Parameters and Esthetic Attributes", Air, Soil and Water Research, 15, p.11786221221075005, 2022.