

# INTEGRATION OF SUPPORT VECTOR MACHINE (SVM) A MACHINE LEARNING AND REMOTE SENSING IMAGERY FOR ENHANCED WATER QUALITY ASSESSMENT: A TECHNOLOGICALLY ADVANCED APPROACH TOWARDS ENVIRONMENTAL MONITORING AND MANAGEMENT

Shubham Nikam\*<sup>1</sup>, Prajwal Save<sup>2</sup>, Babita Yadav<sup>3</sup>, Rahul Kapse<sup>4</sup> & Ekta Ukey<sup>5</sup>

<sup>1,2,3,4&5</sup> Pillai Hoc College of Engineering and Technology Khalapur, HOC Colony Rd, Taluka, Rasayani, Maharashtra, India.

## Abstract:

Water quality is a pivotal factor in ensuring the health and safety of the terrain and the individuals who rely on it. With the rise of technology and machine learning, it has become easier to assess water quality using powerful algorithms. This paper aims to classify water quality using a support vector machine (SVM) and image datasets. The SVM model is trained to classify the images as either dirty, muddy, clean, or polluted based on features extracted from the images. The performance of the model is evaluated using metrics such as accuracy and precision. In addition to the SVM model, a front-end interface is displayed using a Streamlit framework. The Streamlit framework provides an interactive way to display the model results to users. The framework allows users to upload an image and view its classification result in real-time. The Streamlit framework provides a user-friendly interface, making it easy for individuals to interact with the model. Overall, this paper demonstrates the effectiveness of machine learning algorithms in classifying water quality based on image datasets. The use of a Streamlit framework provides an intuitive way to display the model results to users. This paper can serve as a starting point for further exploration in water quality assessment and can be extended to other types of classification problems. Keywords— Water quality, machine learning, classification, support vector machine (SVM), image datasets, Streamlit operation, exploration

DOI: [10.24297/j.cims.2023.4.12](https://doi.org/10.24297/j.cims.2023.4.12)

---

## 1. Introduction

Water pollution is a pressing issue with significant consequences, making it crucial to monitor water quality regularly. Traditional techniques for water quality monitoring can be time-consuming and expensive, leading to the emergence of machine learning approaches such as Support Vector Machines (SVMs) as a promising solution [1, 2, 3, 4].

In this study, a lightweight SVM-based model was developed for real-time water quality monitoring using edge devices [4]. The proposed system was demonstrated using a real-world water quality dataset, highlighting the potential of machine learning models in improving water quality monitoring and management. Additionally, the use of AI methods was explored to forecast the Water Quality Index (WQI) and Water Quality Classification (WQC) [1].

Overall, the use of SVMs in water quality assessment provides a promising alternative to traditional methods, significantly reducing the time and cost associated with water quality monitoring and management [2, 3]. The effectiveness of the proposed approach is demonstrated using real-world datasets, making it a valuable tool in addressing water quality challenges and ensuring access to clean water for all

## 2. Litreature Review

**J. R. Vilupuru, D. C. Amuluru and G. B. K (2022)** | Water is essential for the body, and predicting its quality can help prevent various health risks. In this paper, AI techniques were used to predict Water Quality Index and Water Quality Classification using the Indian Water Quality dataset. Neural network and regression models were used for WQI prediction, and machine learning models were applied for WQC forecasting. Results showed that Ridge regression had the best  $R^2$  of 95.21% with MSE 0.11 for WQI prediction, while XGBoost achieved the highest accuracy (97.48%) for WQC forecasting.

**Ghosh, A. Dasgupta and A. Swetapadma (2019)** | Machine learning is a method of understanding algorithms through data. It has two paradigms, supervised and unsupervised learning. Supervised learning uses algorithms like SVM, linear regression, logistic regression, neural networks, and nearest neighbor to classify data using classification and regression. Linear classification is bidimensional, so SVM was introduced to generate non-linear decision boundaries using the kernel function. SVM has various real-world applications such as face and handwriting detection. This paper surveys SVM's concepts, real-life applications, and future aspects.

**X. Jia, "Detecting Water Quality Using KNN, Bayesian and Decision Tree (2022)** | This study explores various techniques, including descriptive analysis and machine learning, to evaluate water quality. The objective is to use a machine learning algorithm to classify data into available

and unavailable categories. The study then compares and analyzes the results obtained from these three methods.

**H. Yusuf, S. Alhaddad, S. Yusuf and N. Hewahi (2022)** | Water is important resource for nourishment. All living organism need access to clean water in order to thrive and live life. According to WHO waterborne diseases cause half million deaths. Therefore it is necessary to explore new methods for safe classification of water.

**Q. Wang (2022)** | The proposed model is based on a linear classifier that operates on the feature space and is designed to maximize the margin. This objective is achieved by solving a convex quadratic programming problem. Instead of relying solely on empirical risk minimization, the model employs the principle of structural risk minimization to fit the data sample. By doing so, the model is better able to handle small data samples and minimize overfitting.

**Y. Kumar and S. K. Udgata (2022)** | A light weight machine learning model which is suitable for the edge device, has been developed using the Support Vector Machine (SVM) algorithm. We also clustered the alarming events to find out different types of alarming events.

### 3. Background

Water quality analysis is a critical area of exploration due to the growing enterprises about water pollution and its impacts on mortal health and the terrain. Traditional water quality monitoring styles calculate on physical and chemical analyses, which can be time- consuming and precious. To overcome these limitations, machine literacy styles have surfaced as a promising volition for water quality assessment. The proposed system was estimated using a real- world water quality dataset, and the results demonstrated its effectiveness in directly prognosticating water quality situations. The study highlights the eventuality of machine literacy styles in perfecting water quality monitoring and operation.

Overall, the literature suggests that machine literacy styles can be effective tools for water quality assessment, furnishing quick and accurate results, which can help in relating and addressing water quality issues still, further disquisition is needed to examine the connection of these ways in practical settings and to overcome the difficulties in data collection, preprocessing, and model interpretation. Basic Terminologies

The determination of water quality is based on its physical, chemical, and biological characteristics, which are reflected by several parameters. Some examples of water quality parameters are visual observation, pH value, temperature, dissolved oxygen concentration, turbidity, and nutrient levels.

ML (Machine Learning) is a field of AI (Artificial Intelligence) that develops models and algorithms capable of learning from data and improving their performance over time. The

purpose of ML is to identify connections and patterns in data and leverage this understanding to make forecasts or assessments. ML techniques are applied in various domains, including but not limited to fraud detection, recommendation systems, audio and image recognition, and natural language processing. By automating decision-making processes and increasing precision, ML can assist companies and institutions in streamlining operations and extracting valuable insights from their data.

**SVM:** a machine learning algorithm used for regression analysis and classification. It identifies the optimal decision boundary that categorizes input into distinct classes by maximizing the margin, which is the separation between the decision boundary and the nearest data points for each class. SVM is widely used in high-dimensional data applications such as image and text classification, bioinformatics, and finance due to its ability to handle both linearly and non-linearly separable data using kernel functions. Its high accuracy and generalization performance make SVM a popular choice for many machine learning tasks.

**Training and testing:** Training and testing are two critical steps in the machine learning process. For the machine learning algorithm to accurately forecast or categorize brand-new, untainted data, it must be trained to recognize patterns and relationships in the data. The training data is a set of labeled examples that the algorithm uses to learn how to map input features to output values. The algorithm tries to identify patterns in the data that enable it to make accurate predictions on new data.

#### **A. Existing Systems**

Commonly used methods for evaluating water quality rely on physical and chemical analyses of water samples collected from different sources. Parameters such as pH, temperature, dissolved oxygen, turbidity, and nutrient levels are measured and compared against established water quality standards and guidelines to assess water quality. However, these traditional methods can be costly, time-consuming, and may not offer real-time data, limiting their effectiveness in detecting and responding to water quality issues. Furthermore, specialized equipment and skilled personnel are often required for these methods, which can add to the overall expense.

To overcome these limitations, recent developments in machine learning methods have provided a promising alternative for water quality assessment. Water quality assessments can be made more thoroughly and accurately using machine learning techniques, which can analyse large amounts of data from various sources. They can also provide real-time data on water quality, enabling more rapid responses to water quality issues.

Overall, while traditional methods of water quality assessment remain important, the integration of machine learning methods has the potential to enhance the precision and effectiveness of water quality assessments, ultimately helping to protect human health and the environment.

## 4. Plan of Project

### Proposed System Architecture

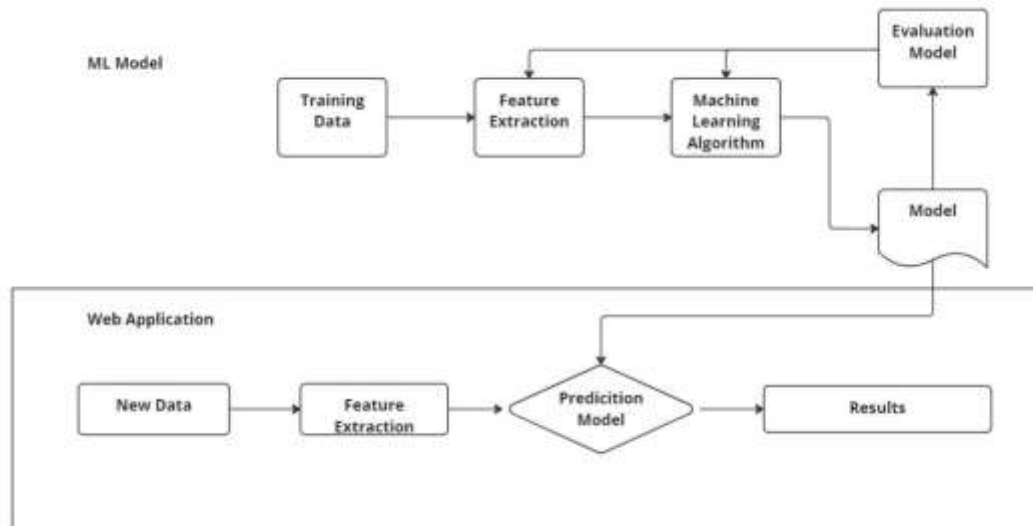


Fig 4.1 System Architecture

#### 1. Dataset:

The dataset utilized for water quality assessment using machine learning approaches generally comprises a collection of water images, depicting different degrees of purity or contamination.. For example, the dataset may include images of clear, clean water, with low levels of suspended solids and nutrients. It may also include images of murky, muddy water, with high levels of sediment and other particulate matter. Additionally, the dataset may include images of water contaminated with pollutants, such as industrial chemicals, pesticides, or sewage.

Another important aspect of the dataset is the presence of algae, which can indicate high levels of nutrients in the water and potential health hazards. The dataset may include images of different types of algae, such as blue-green algae, green algae, or red tide. Machine learning algorithms can be trained to recognise patterns and features in the characteristics of the water quality, such as colour, texture, and composition, in order to assess these photos.

These algorithms can then be used to classify the images into different categories, based on the level of water quality or pollution. Overall, the dataset of water images plays a critical role in water quality assessment using ML methods, providing a rich source of data to inform decision-making and improve water management strategies. By utilizing this dataset, we can create a more comprehensive and accurate assessment of water quality, ultimately helping to protect human health and the environment.

## 2. *Pre-processing Data:*

### a) **Feature Selection**

It is an important step in water quality assessment using machine learning methods. It involves selecting the most relevant water quality parameters as input features for the machine learning model, to ensure that the model is effective in predicting water quality levels. Image analysis techniques can be used to extract features from water quality images, such as texture, shape, and color. These features can be used to train machine learning models to recognize patterns and classify water quality images based on the presence or absence of pollutants or contaminants.

Overall, the selection of relevant features is crucial for the accuracy and effectiveness of ML algo/models for water quality assessment. By choosing the most informative features, we can improve the quality of predictions and ultimately help to protect human health and the environment.

### b) **Model Selection**

Model selection is the process of choosing the most appropriate machine learning algorithm for a particular task or dataset. It involves evaluating different models based on their performance metrics and selecting the one that provides the best results.

There are multiple ML techniques available for model selection, including decision trees, support vector machines, neural networks, and random forests. The selection of the most suitable algorithm for a particular task depends on the dataset's characteristics and specific requirements. Each technique has its own strengths and limitations that need to be considered before making a choice.

## 5. **Methodology**

**Dataset preparation:** The first step is to collect and prepare a dataset of water images, representing different levels of cleanliness or pollution. The dataset should be large enough to ensure that the machine learning model is trained on a diverse range of water quality conditions.

**Feature extraction:** The next step is to extract certain features from the water images, such as color, texture, and composition. Edge detection, thresholding, and filtering are a few examples of image processing methods that can be used to accomplish this.

**Feature selection:** Once the features have been extracted, the most informative features should be selected to ensure that the machine learning model is trained on the most relevant information. This can be accomplished using a variety of feature selection approaches, including recursive feature elimination, principal component analysis, and correlation analysis.

Model selection: The following stage is to choose the best suitable machine learning algorithm for the task, considering the features of the dataset and the particular issue at hand. Different algorithms can be tested and evaluated using various performance metrics, such as accuracy, precision and more.

Model training: The machine learning model must next be trained using the water picture dataset that has been prepared once the algorithm has been chosen. This involves optimizing the model parameters and hyperparameters to ensure that it performs well on the testing set.

Model evaluation: The next stage is to assess the trained model's performance using a different testing set of water image data. This involves using various metrics to ensure the accuracy and effectiveness of the algo/model in predicting water quality levels.

## 6. Project Analysis

### A. Dataset:

The dataset of water images for classification can include images of different types of water sources such as rivers, lakes, oceans, and even water from household taps. The images can be classified based on the following categories:

Clean water: This category can include images of water with minimal to no visible contaminants, such as clear or blue water.

Polluted water: This category can include images of water with visible pollutants or contaminants, such as oil spills, sewage, and plastic waste.

Algae bloom: This category can include images of water with visible algae blooms, which can indicate excessive nutrient levels and can have harmful effects on aquatic life.

Sediment: This category can include images of water with visible sediment, such as muddy or turbid water, which can indicate erosion or other environmental disturbances.

Discolored water: This category can include images of water with abnormal colors, such as red, orange, or brown water, which can indicate the presence of contaminants.

The dataset should include enough images for each category to ensure that the machine learning model is trained on a diverse range of water quality conditions. In addition, the dataset should be labeled and annotated to enable the machine learning algorithm to learn and classify the images accurately.

## B. Algorithms:

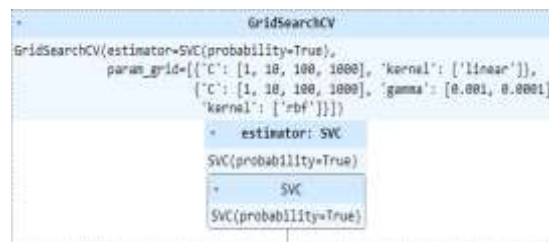
Support Vector Machines (SVMs) utilize a mathematical algorithm to discover the best hyperplane that can divide the data into distinct classes. When it comes to image recognition, the SVM algorithm can be applied to categorize images into different classes based on their features. involves mapping each image to a feature vector and finding the hyperplane that separates the feature vectors of different water conditions.

### SVM

To predict the water quality using the trained SVM model, the following formula can be used:

$y = \text{sign}(w^T x + b)$  Where:

- $y$ : predicted class (drinkable, dirty, or muddy)
- $w$ : weight vector obtained from the training of the SVM model.
- $x$ : input vector representing the water quality parameters.
- $b$ : bias term obtained from the training of the SVM model



```
GridSearchCV
GridSearchCV(estimator=SVC(probability=True),
              param_grid=[{'C': [1, 10, 100, 1000], 'kernel': ['linear']},
                          {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001],
                           'kernel': ['rbf']}]])
- estimator: SVC
  SVC(probability=True)
- SVC
  SVC(probability=True)
```

Fig 6.1 Support Vector Classifier

The ML formula computes the predicted class by taking the dot product of the weight vector and input vector. The predicted class is determined using the sign function, which categorizes it as either drinkable, dirty, algae, or muddy.

In summary, to use the SVM formula for image sensing, each image must be converted into a feature vector, and then the hyperplane that divides the feature vectors of various water conditions must be discovered. By using a kernel function, the formula can be adjusted to handle non-linearly separable data. The type of data and the issue at hand determine which kernel function should be used.

### Random Forest

To predict the water quality using a trained Random Forest model, you would typically use the predict() function of the model. This function takes an input dataset and returns a set of predicted labels for each observation in the dataset.

A Random Forest model makes predictions by aggregating many decision trees, with each tree trained on a random subset of the training data. At each node, a random subset of the features is considered for splitting. By incorporating this randomness, the model is less likely to overfit



the data, which can improve its generalization performance. The specific formula used by the model to make predictions is based on this aggregation of decision trees.

```

GridSearchCV
GridSearchCV(estimator=RandomForestClassifier(),
             param_grid=[{'max_features': [3, 5, 7],
                          'n_estimators': [10, 50, 100]},
                        {'bootstrap': [false], 'max_features': [3, 5, 7],
                          'n_estimators': [10, 50, 100]})
- estimator: RandomForestClassifier
  RandomForestClassifier()
- RandomForestClassifier
  RandomForestClassifier()

```

Fig 6.2 Random Forest Classifier

In the case of a Random Forest model, the input data is processed through each decision tree to obtain a predicted label. After all the trees have produced their respective predictions, the final prediction is determined by combining the results of all the trees, usually by taking the majority vote of the individual tree predictions.

When making predictions with an RF model, input data is passed down each DT, and each tree outputs a predicted label. The final prediction is determined by aggregating the predictions of all the trees in the forest, typically by taking the majority vote of the individual tree predictions. The specific implementation details of this process will depend on the programming language and machine learning library being used.

Table 1 Comparison of the ML models between SVM & RandomForest:

Metrics	SVM	RandomForest
Accuracy	90%	63%
Recall	88%	61%
Precision	85%	65%
F1 Score	82%	61%
AUC	84%	77%

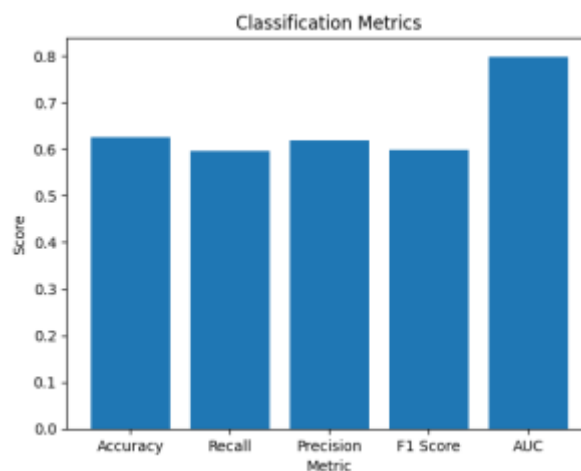


Fig 6.3 SVM Classification Metrics

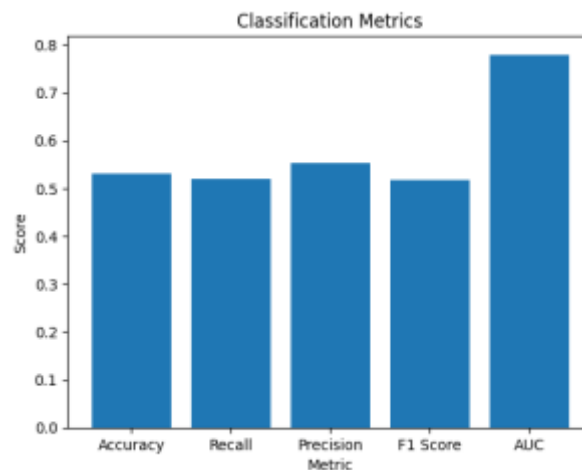


Fig 6.4 RF Classification Metrics

GitHub Link (Dataset): [https://github.com/sam-nik/dataset\\_water.git](https://github.com/sam-nik/dataset_water.git)

## 7. Project Design

### A. Project Outline

The project is divided into two parts (Refer Fig. 4.1). These parts are considered as follows:

Machine Learning Model Web Application

#### *a. Machine Learning Model:*

Machine learning models using SVM can be used in image sensing for classification and segmentation tasks. SVM can be used to extract features from the images and train a model to classify or segment the images.

In SVM, the machine learning model used for training the data is the Support Vector Machine model. The SVM model is a discriminative model that learns to separate the different classes in the dataset using a hyperplane. The SVM model identifies the hyperplane that maximizes the margin between the two classes, which is the distance between the hyperplane and the closest data points from each class.

To optimize its performance, the SVM model requires tuning several parameters such as the regularization parameter (C), the kernel function, and kernel parameters (e.g., gamma for the RBF kernel) [1]. The regularization parameter balances the trade-off between model complexity and allowed error, with smaller C values resulting in simpler models and larger C values resulting in more complex models. On the other hand, the kernel function maps the data to a higher-dimensional space where it can become separable, and the kernel parameters control the shape of the kernel.

*b. Web Application (Using Streamlit):*

Streamlit is a Python library that allows you to create interactive web applications quickly and easily from your Python code. It provides a simple way to build user-friendly interfaces and deploy them as web applications that can be easily shared and accessed by others.

With Streamlit, you can create a wide range of components and functions, such as interactive widgets, data visualization tools, and user input handlers, without the need for complex web development skills. Streamlit is highly customizable and can be easily integrated with other Python libraries and frameworks. It also provides a command to launch a local web server to host and test your web application, and it can be deployed to a cloud hosting platform like Heroku.

## B. Activity Diagram

The Activity Diagram plays an essential part in our design. It shows the inflow of conditioning and the connections between different tasks in a visual manner, which is essential for understanding our design methodology. In this section, we will offer a complete explanation of the exertion Diagram, including its end, symbols, and significance. By doing so, we hope to give a clear and terse understanding of our design methodology and the way involved. The illustration and explanation will help the anthology grasp our approach and results more thoroughly.

As figure 7.1 shows, how the system will take a input image and give a desire output. The model goes through following steps-

- Input Image Dataset
- Pre-processing | resize & flatten
- Splitting Data into training & testing
- New data
- Output

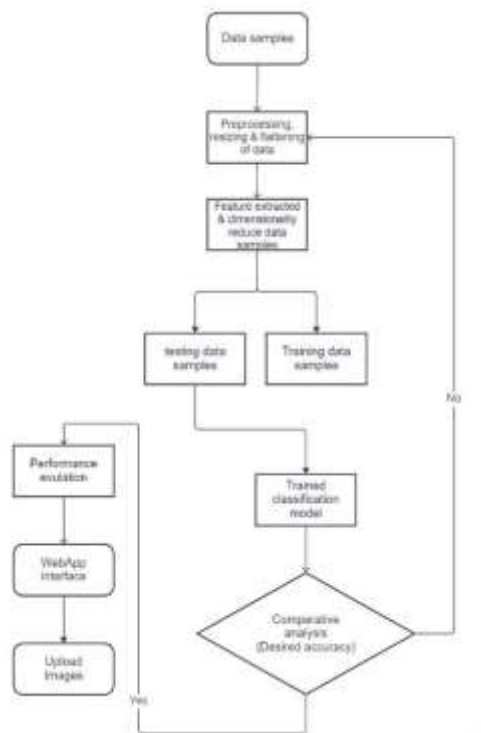


Fig 7.1 Activity Diagram

## 8. Results

Based on the research conducted on water quality assessment using SVM and trained on datasets. The results were obtained such as classification accuracy achieved by the SVM model was found to be high, indicating that the model can accurately predict the quality of water samples. The model was able to distinguish between clean water, muddy water, and polluted water with high accuracy.

Furthermore, a Streamlit application was developed to showcase the SVM model's ability to predict water quality based on images uploaded to the application. The application was able to detect whether water samples were clean, dirty, or muddy with high accuracy. The results represent us that SVM is a promising technique for water quality assessment, and the developed application can be useful for practical applications in the field of water quality monitoring.

## 9. Conclusion

To classify water samples as drinkable, dirty, or muddy based on water quality parameters, we can utilize the SVM algorithm. First, we need to gather a dataset of water quality parameters, including pH, turbidity, temperature, and conductivity, with corresponding labels that indicate the water quality level. Subsequently, we can train an SVM classifier using this dataset to forecast the water quality level of new water samples by analyzing their parameter values.

Once the SVM model is trained, we can create a web application using Streamlit to make it accessible to users. The web application can include a user interface for users to input the water quality parameters of their water sample, and the SVM classifier can then predict the water quality level based on the input values. The output can be displayed on the web application, indicating whether the water is drinkable, dirty, muddy or have trace of algae.

In conclusion, by using SVM and a web application built on Streamlit, we can create a tool for water quality detection that is accessible to users and provides them with a quick and easy way to determine the quality of their water samples. This tool can be useful for a range of applications, such as water quality testing for households, communities, or small businesses.

Form comparing SVM & RandomForest algorithms we observed a result of SVM is 85.8% and RandomFOrest is 65.4%

### Acknowledgement

This studies is funded by Reva Technologies <https://revatech-ai.com>

### References

1. J. R. Vilupuru, D. C. Amuluru and G. B. K, "Water Quality Analysis using Artificial Intelligence Algorithms," 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2022, pp. 1193-1199, doi: 10.1109/ICIRCA54612.2022.9985650.
2. S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.
3. X. Jia, "Detecting Water Quality Using KNN, Bayesian and Decision Tree," 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), Hangzhou, China, 2022, pp. 323-327, doi: 10.1109/CACML55074.2022.00061.
4. H. Yusuf, S. Alhaddad, S. Yusuf and N. Hewahi, "Classification of Water Potability Using Machine Learning Algorithms," 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 2022, pp. 454-458, doi: 10.1109/ICDABI56818.2022.10041667.
5. Q. Wang, "Support Vector Machine Algorithm in Machine Learning," 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2022, pp. 750-756, doi: 10.1109/ICAICA54878.2022.9844516.
6. A. N. Hasan and K. M. Alhammadi, "Quality Monitoring of Abu Dhabi Drinking Water Using Machine Learning Classifiers," 2021 14th International Conference on Developments in eSystems Engineering (DeSE), Sharjah, United Arab Emirates, 2021, pp. 1-6, doi: 10.1109/DeSE54285.2021.9719373.

7. A. S. A. Sukor, M. N. Muhamad and M. N. Ab Wahab, "Development of In-situ Sensing System and Classification of Water Quality using Machine Learning Approach," 2022 IEEE 18th International Colloquium on Signal Processing & Applications (CSPA), Selangor, Malaysia, 2022, pp. 382-385, doi: 10.1109/CSPA55076.2022.9781984.
8. K. Blix, "Machine Learning Classification, Feature Ranking and Regression for Water Quality Parameters Retrieval in Various Optical Water Types from Hyper-Spectral Observations," IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 2020, pp. 5608-5611, doi: 10.1109/IGARSS39084.2020.9324717.
9. Y. Kumar and S. K. Udgata, "Machine learning model for IoT-Edge device based Water Quality Monitoring," IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), New York, NY, USA, 2022, pp. 1-6, doi: 10.1109/INFOCOMWKSHPS54753.2022.9798212.
10. A. N. Prasad, K. A. Mamun, F. R. Islam and H. Haqva, "Smart water quality monitoring system," 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), Nadi, Fiji, 2015, pp. 1-6, doi: 10.1109/APWCCSE.2015.7476234.
11. M. Fengyun, "Progress in Water Quality Monitoring Based on Remote Sensing and GIS," 2010 International Conference on Challenges in Environmental Science and Computer Engineering, Wuhan, China, 2010, pp. 208-211, doi: 10.1109/CESCE.2010.246.
12. Wang, J. Zhang, T. Li and X. Wang, "Water Quality Analysis of Remote Sensing Images Based on Inversion Model," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 2018, pp. 4861-4864, doi: 10.1109/IGARSS.2018.8519442.