International Journal of Computational Science and Engineering.
ISSN 2249-4251 Volume 13, Number 1 (2023), pp. 45-55
© Research India Publications
https://dx.doi.org/10.37622/IJCSE/13.1.2023.45-55

Water Quality Prediction using Image Processing and Machine Learning Models

Mrs. Jayashree S Patil[1], Akhila Mailaram[2], Pavani Naga Kumari Basa[3], A. Sai Sravya[4] and Banvita Yadam[5]

G. Narayanamma Institute of Technology and Science (for Women)
Shaikpet, Hyderabad, Telangana, India.
[1]Associate Professor, [1]jshivshetty@gnits.ac.in, [2]akhilamailaram0312@gmail.com, [3]pavsai06.basa@gmail.com, [4]sravyaaman@gmail.com, [5]banvitayadam@gmail.com

Abstract:

The quality of water pertains to its physical, chemical, biological, and sensory properties that determine its appropriateness for various applications. Assessing and determining water quality is crucial as it directly affects human health, industrial and domestic purposes, and the natural ecosystem. Water quality can be evaluated through laboratory techniques or test kits designed for home use. Although laboratory analysis provides the most precise results, it is also time-consuming. The water quality parameters tested must conform to the standards set by local authorities, which are often influenced by international regulations created by the water quality organizations such as the World Health Organization (WHO). As a result, this study proposes a water quality detection system using machine learning methods. The crucial features of water are extracted from water images using machine learning methods based on color features. The proposed system presents comparison of accuracies of different machine learning models. The highest accuracy is achieved through multi-layer perceptron predicted to be 93.92% on feature dataset.

Keywords: Machine Learning, Color Moments, Multilayer Perceptron, Features

1. Introduction

Have you ever wondered about how we obtain insights on the state of our country's waterways, including lakes, streams, coastal waters, and estuaries? Monitoring is the primary means of obtaining this data. There are numerous methods to keep track of state of water. Experts in monitoring examine the chemical composition of water, deposits, and aquatic organisms like fish tissue to measure concentrations of key components, such as nutrients, metals, pesticides, dissolved oxygen and oils. They also

observe environmental factors, such as temperature, rate of flow, sediment composition, and the potential for erosion along lake shores and stream banks. Increasingly, monitoring efforts are being directed towards assessing the state of watersheds or hydrological basins, which are the areas drained by lakes, estuaries, and lakes [7]. Currently, the utilization of digital image processing techniques is gaining popularity as it provides a low-cost, visual, and contactless approach [10].

Numerous studies have previously focused on image recognition, [14] where a specific classifier such as Machine Learning is used to categorize images into different classes. This technology has been implemented in many fields, including plant disease detection [1], medical image processing, food analysis, intelligent manufacturing and biometrics and has achieved outstanding results. Feature extraction and image classification are the key processes in machine learning for image recognition, and the final result of image recognition directly depends on effectiveness of feature extraction.

This study presents the extraction of essential characteristics from water images based on color. The extracted features include the color moments of each RGB channel, which are then utilized for evaluating water quality through image classification. By merging the strengths of various machine learning models, the classification of water quality images is carried out.

2. Related Work

Water quality is a global problem and many international institutions, scientists, and individuals work together to effectively monitor the work conditions. There are many methods through which water quality is monitored. In-situ sensors, [3] These are devices that are deployed directly in the water to measure various water quality factors, such as temperature, pH, turbidity, dissolved oxygen, and conductivity, laboratory analysis where water samples are collected from the site and transported to a laboratory for analysis. Common laboratory techniques include colorimetry, titration, and spectrophotometry, remote sensing [9] through satellite or aerial imagery is used to monitor the water quality factors, such as chlorophyll-a concentration, turbidity, and [13] surface temperature, bio-indicators [5] in which biological organisms, such as algae and macroinvertebrates, are used as indicators of water quality, changes in the abundance and diversity of these organisms can signal changes in water quality. Chemical sensors [8] are devices that use chemical reactions to detect and measure specific contaminants in the water, such as heavy metals, pesticides, and organic compounds, online monitoring [12] is real-time data collection from various sensors and transmitted to a central location for processing and analysis. This allows for continuous monitoring of water quality parameters, there have been many contributions made through online monitoring using [2] big data, [4] artificial intelligence, [6] deep learning and so on, citizen science is where community members are trained to collect water samples and use simple testing kits to evaluate basic water quality metrics, such as pH and temperature through organisations like WHO or [11] United Nations water protection agency.

3. Methodology

3.1 Data Acquisition

Around 2100 pictures are gathered from well-known websites such as Shutter stock, Google earth engine, and others. These sample images are categorized into distinct classes based on internet knowledge and converted to a standard jpg format. Each type comprises a specific set of images, with a whole of four classes for the collected samples that represent varying degrees of water quality. Specifically, these four classes are algae water, clean water, mud water, and polluted water. To facilitate subsequent computations, these pictures are converted into the RGB model and then resized to 150×150 pixels to fit the model.

3.2 Overview

Illustrated in Figure 1, the entire methodology for detection of water quality is outlined. Initially, appropriate images of water samples are gathered. These sample images are categorized based on internet knowledge and labelled to construct a sample library for modelling purposes. Concurrently, we create feature engineering by extracting and selecting key features such as the color moments of each color channel to build the model. Subsequently, the model is trained by feeding the feature samples into the machine learning algorithms proposed in our approach.

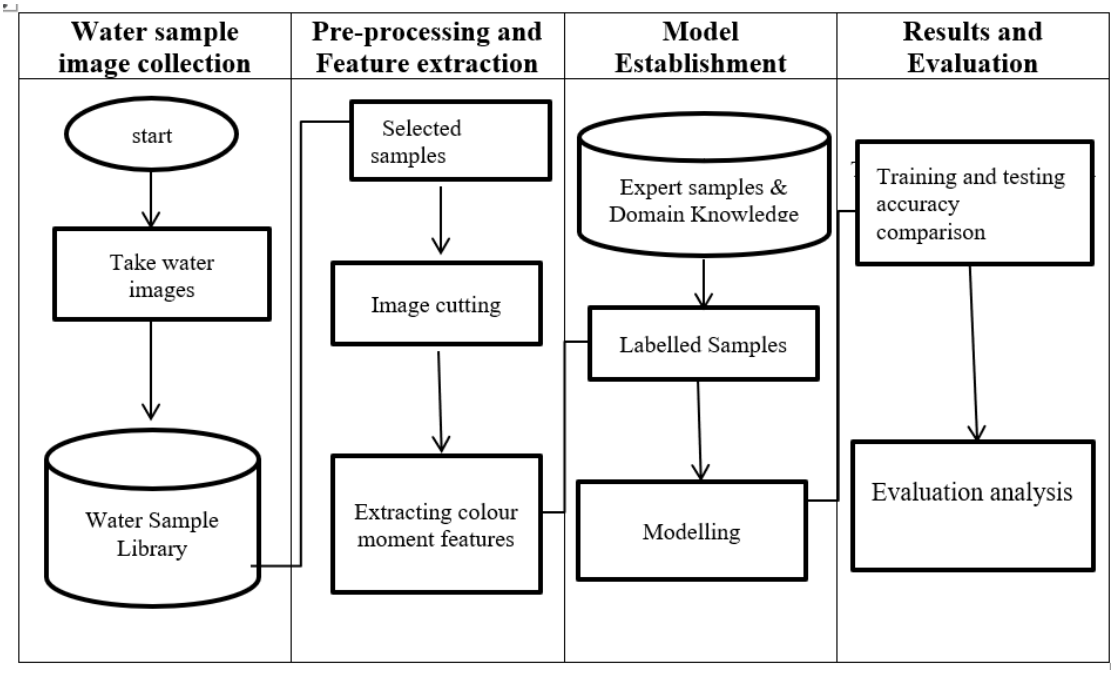

Fig. 1. Flow chart of the proposed approach

4. Proposed Approach

4.1 Image Processing

Since the images of the collected water samples from the website contain additional

elements such as humans, boats, and birds, it is imperative to pre-process them beforehand. The primary area of interest in the sampled images is the water body, and therefore, the incision of image technique can be employed to obtain the water's color features by cutting out a portion of the sample image. A sub-image of 150 × 150 pixels can be cut from the center of the collected sample images for feature extraction. By using this method, the water features can be obtained from the sub-images, allowing for further data processing.

4.2 Feature Extraction

The recognition of images through machine learning involves two key processes, feature extraction and then image classification. Of these, feature or attribute extraction holds greater importance since the quality of extracted features directly impacts image prediction outcomes. As is commonly known, image features encompass a range of characteristics, such as color, texture, shape, spatial relationships, and more. Color features are particularly noteworthy due to their stability, insensitivity to object size and direction, and high level of robustness. For instance, in this system, where water images exhibit uniformity, the focus is largely on color features, specifically, color moments that are chosen for this study. The study describes the extraction of the first order, second order, and third order moments of RGB color channels, with detailed calculation procedures outlined below.

First order moment:

The initial color instance may be construed as the mean color within the picture and computed by employing the subsequent equation 1

$$M_{i1} = \frac{1}{N}\sum_{j=1}^{N} Pij \quad (1)$$

Here, i represents a component of the RGB color channels, N denotes the total number of pixels in the image, and Pij refers to the value of the jth pixel in color channel i.

Second order moment:

The next moment of color is the standard deviation, which can be calculated by computing the square root of the variance present in the color distribution as shown in equation 2.

$$M_{i2} = \sqrt[2]{\frac{1}{N}\sum_{j=1}^{N}(Pij - Mi1)\text{^}2} \quad (2)$$

where Mi1 denotes the first order moment of the ith channel, Mi2 denotes the second order moment of the ith channel.

Third order moment:

The third chromatic moment is known as skewness. It evaluates the degree of asymmetry in the color distribution, providing insight into its shape. Skewness can be calculated using the subsequent equation.

$$\text{Mi3} = 3\sqrt{\frac{1}{N}\sum_{j=1}^{N}(Pij - Mi1)\text{^}3} \quad (3)$$

where Mi3 denotes the third order moment of the ith channel.

The color moments obtained are tabulated and labelled with the respective classes and resulted data set formed is as shown in the figure table 1.

Table 1. Table depicting the csv file obtained by extracting color moments

	A	B	C	D	E	F	G	H	I	J	K
1	Image_Name	First_order_moment			Second_order_moment			Third_order_moment			class
2		R	G	B	R	G	B	R	G	B	
3	'Clean_water\\1 (2).jpg'	0.0185	1.2714	1.3673	0.0255	0.1517	0.1101	0.0176	-0.0026	-0.0002	Clean
4	'Clean_water\\1 (3).jpg'	0.0088	0.7979	1.2789	0.0116	0.0633	0.0586	0.0152	-0.001	-0.0084	Clean
5	'Clean_water\\1 (4).jpg'	0.3818	1.7833	1.9392	0.6184	0.4045	0.214	0.0134	-0.0166	-0.0173	Clean
6	'Clean_water\\1 (5).jpg'	1.1441	1.6544	1.9084	0.6948	0.568	0.5013	-0.0008	-0.0042	-0.0073	Clean
7	'Clean_water\\1 (7).jpg'	1.2278	1.9341	1.5803	0.4394	0.1917	0.249	-0.0016	-0.0059	-0.0044	Clean
8	'Clean_water\\10 (2).jpg'	0.8745	1.9625	2.1816	0.3721	0.1701	0.1334	0.0106	0.0084	0.0055	Clean
9	'Clean_water\\10 (3).jpg'	0.7155	1.8573	1.8753	0.2746	0.0837	0.0929	0.0015	0.0011	0.0029	Clean
10	'Clean_water\\10 (5).jpg'	0.8795	1.743	1.8713	0.1008	0.0884	0.0801	-0.0011	-0.0177	-0.0213	Clean
11	'Clean_water\\10 (6).jpg'	1.6817	1.9769	1.8847	0.1592	0.101	0.1178	-0.0061	-0.0016	-0.0011	Clean
12	'Clean_water\\10 (7).jpg'	1.63	2.1153	2.1463	0.3688	0.1529	0.1652	0.0038	0.0072	0.0058	Clean
13	'Clean_water\\10 (8).jpg'	1.6702	1.9746	1.8857	0.1839	0.1227	0.1376	-0.0074	-0.0023	-0.002	Clean
14	'Clean_water\\100 (2).jpg'	0.1511	1.4782	1.5216	0.1029	0.1296	0.1232	0.0051	-0.0007	0.0001	Clean
15	'Clean_water\\100 (3).jpg'	0.6144	2.2204	2.3376	0.3194	0.1446	0.1052	0.0016	-0.001	0.0003	Clean
16	'Clean_water\\101 (2).jpg'	0.0074	2.0012	1.6974	0.0324	0.0778	0.0569	0.0292	-0.0037	0.0012	Clean
17	'Clean_water\\102 (2).jpg'	0.4324	1.42	1.3554	0.0356	0.0848	0.0804	0.0024	0.0001	-0.0006	Clean
18	'Clean_water\\102 (3).jpg'	1.334	2.1918	2.2933	0.2425	0.0498	0.053	-0.0034	0.0035	0.0054	Clean
19	'Clean_water\\102 (4).jpg'	1.3934	1.9673	2.0447	0.282	0.1684	0.0997	-0.0066	-0.0124	-0.0097	Clean
20	'Clean_water\\102 (5).jpg'	0.0378	1.2181	1.7718	0.1441	0.3207	0.2299	0.0435	0.0078	0.005	Clean
21	'Clean_water\\102 (6).jpg'	0.4197	1.395	1.3235	0.0355	0.1035	0.0822	0.0016	0.002	0.0007	Clean
22	'Clean_water\\103 (2).jpg'	0.0122	1.6001	1.6266	0.0314	0.1907	0.1462	0.0277	0.0008	0.0018	Clean
23	'Clean_water\\103 (3).jpg'	0.9695	1.9926	1.9189	0.3208	0.0704	0.0715	-0.0059	-0.004	-0.0025	Clean
24	'Clean_water\\103 (4).jpg'	1.1855	1.9555	2.1273	0.3668	0.1238	0.0533	-0.0012	-0.0038	-0.0029	Clean
25	'Clean_water\\103 (6).jpg'	0.007	1.272	1.3557	0.0088	0.1739	0.1614	0.0122	0.005	0.0063	Clean

4.3 Machine Learning Models used

To solve the image classification problems for diverse sample images, various classifiers are combined with underlying features extracted from the images. Popular machine learning algorithms in image recognition include K Nearest Neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Networks (ANNs), Decision Trees and Random Forests (RF).

SVM is a powerful algorithm that finds the best boundary to separate two classes by maximizing the margin between them. KNN, on the other hand, is a simple algorithm that classifies new instances by finding the k closest neighbors in the training set and assigning the class that is most common among them. Random Forest is an ensemble algorithm that creates multiple decision trees and combines their predictions to make a final decision. Decision Trees are simple and interpretable algorithms that split the data based on the most important features and create a tree-like structure to classify new instances. They are prone to overfitting, but pruning techniques can help reduce this

issue. Naive Bayes Classifier is a probabilistic algorithm that uses Bayes' theorem to calculate the probability of a class given the features of a new instance. Finally an MLP, which stands for Multi-layer perceptron, is a type of ANN or artificial neural network employed for classification and regression purposes. As shown in figure 2, it comprises of an input layer, one or more hidden layers, and an output layer. In the input layer, the nodes represent input variables, while the nodes in the output layer indicate the output variables. The hidden layers consist of nodes that use nonlinear transformations to process the input variables and transfer the output to the subsequent layer. During the training process, the MLP modifies the weights of the connections between the nodes to minimize the discrepancy between the expected and actual output. MLPs are versatile and capable of understanding complex nonlinear correlations between input and output variables; however, they can be vulnerable to overfitting and need vast amounts of training data.

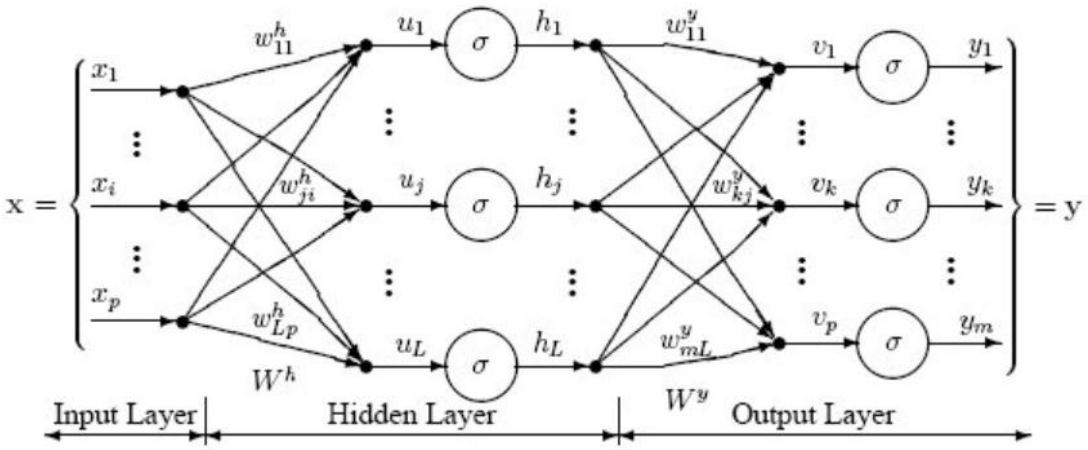

Fig. 2. Internal working of MLP

5. Results and Discussions

By utilizing the approach specified in Section 4.1, the images are sliced, and the recorded sub images of water sample images are introduced. Subsequently, the pre-processing procedures are executed, and each sliced sub-image is categorized by class label and index number. Then, using the technique described in Section 4.2, the color moment characteristics for all channels are derived. A portion of the extracted water quality image features is exhibited in Table 1. Afterward, the feature dataset is randomized, and various machine models are employed. The data is partitioned into training and testing sets with an 8:2 ratio. Numerous experiments are carried out on the dataset using the randomized training data, and the best model is selected for the prediction of the respective class of water images. The outcome evaluation also takes into account the metrics of true negatives and false positives. Hence, in addition to the accuracy, the precision, recall and f1-score have also been considered to assess the models' performance, as represented in the subsequent equations 4, 5 6 and 7 :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (7)$$

where TN indicates the correct negatives, FP indicates the false positives, FN indicates false negatives and TP indicates the true positives respectively.

As demonstrated in Section 4.3, the suggested techniques of machine learning are utilized to carry out both the training and testing of the feature dataset model. The above mentioned metrics are utilized to confirm the efficiency of the algorithms. The dataset is partitioned into a training set and a test set in the ratio of 8:2. This means that approximately 80% of the samples are randomly selected as the training data, while the remaining 20% is assigned to the test set. Well known algorithms as mentioned in section 4.4 are selected for the comparative analysis. Fig 5.1 shows the variations in the accuracy of Decision Tree algorithm based on the depth, where as Fig 5.2 shows the variation in accuracies generated by Random Forest classifier based on number of estimators used. Fig 5.3 depicts the varied accuracies of KNN algorithm based on number of neighbours, and Fig 5.4 depicts the confusion matrix of MLP classifier. The classification accuracies of different methodologies on testing datasets are depicted in Table. 5.1.

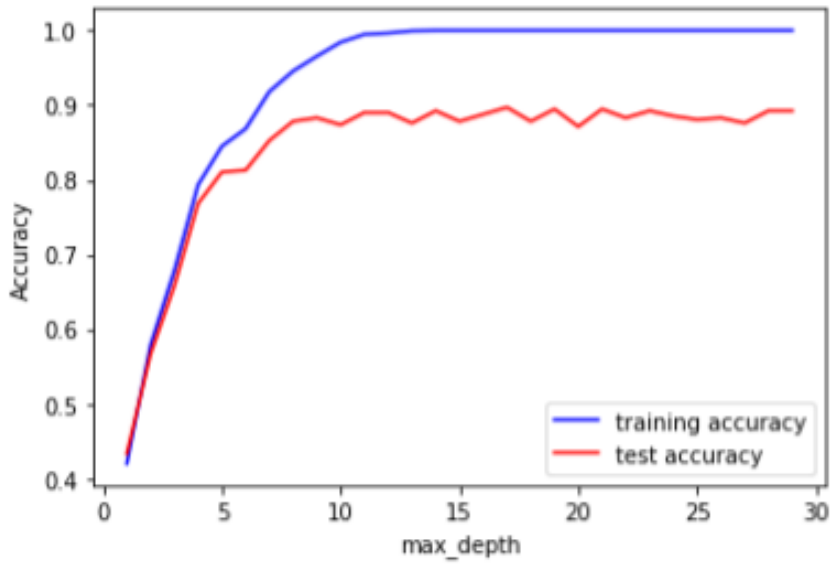

Fig. 3. Decision tree metrics based on depth

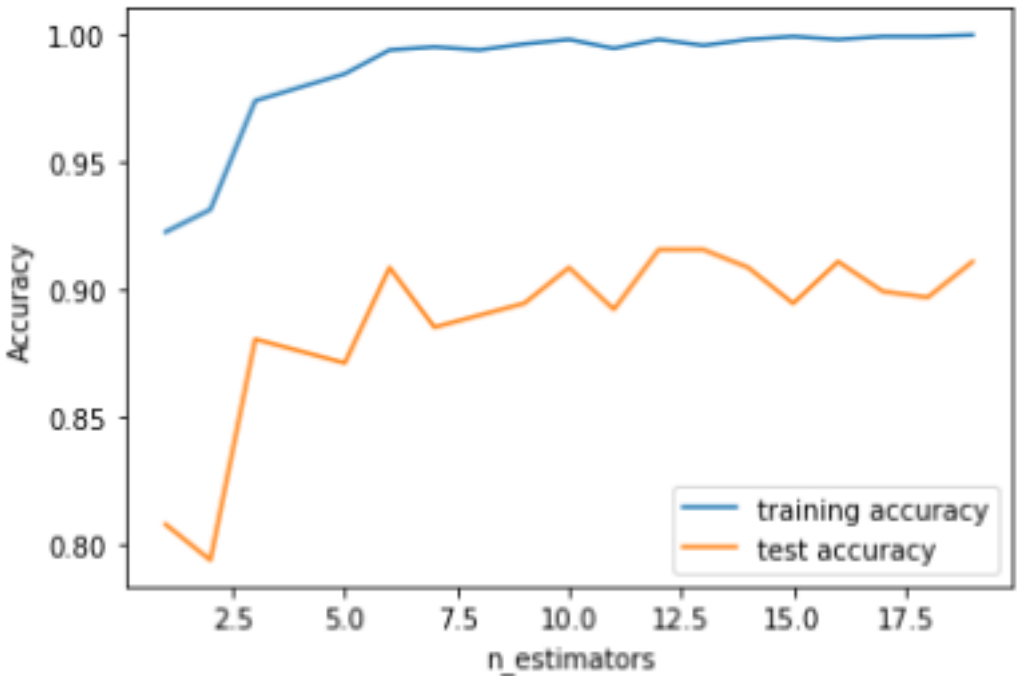

Fig. 4. RF metrics based on no.of estimators

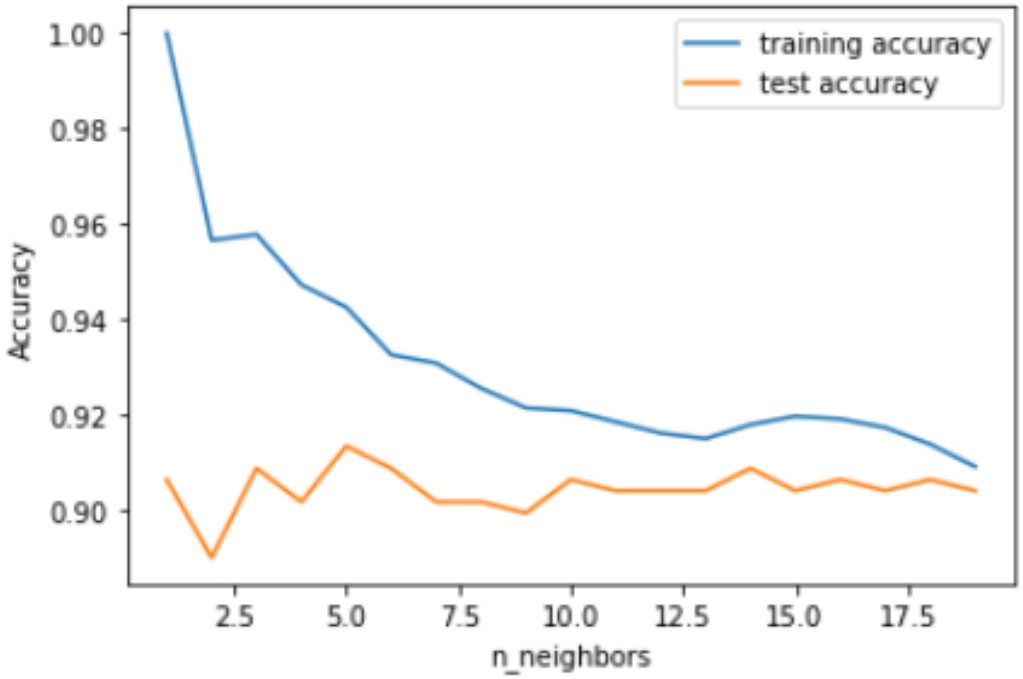

Fig. 4. KNN metrics based on no.of neighbours

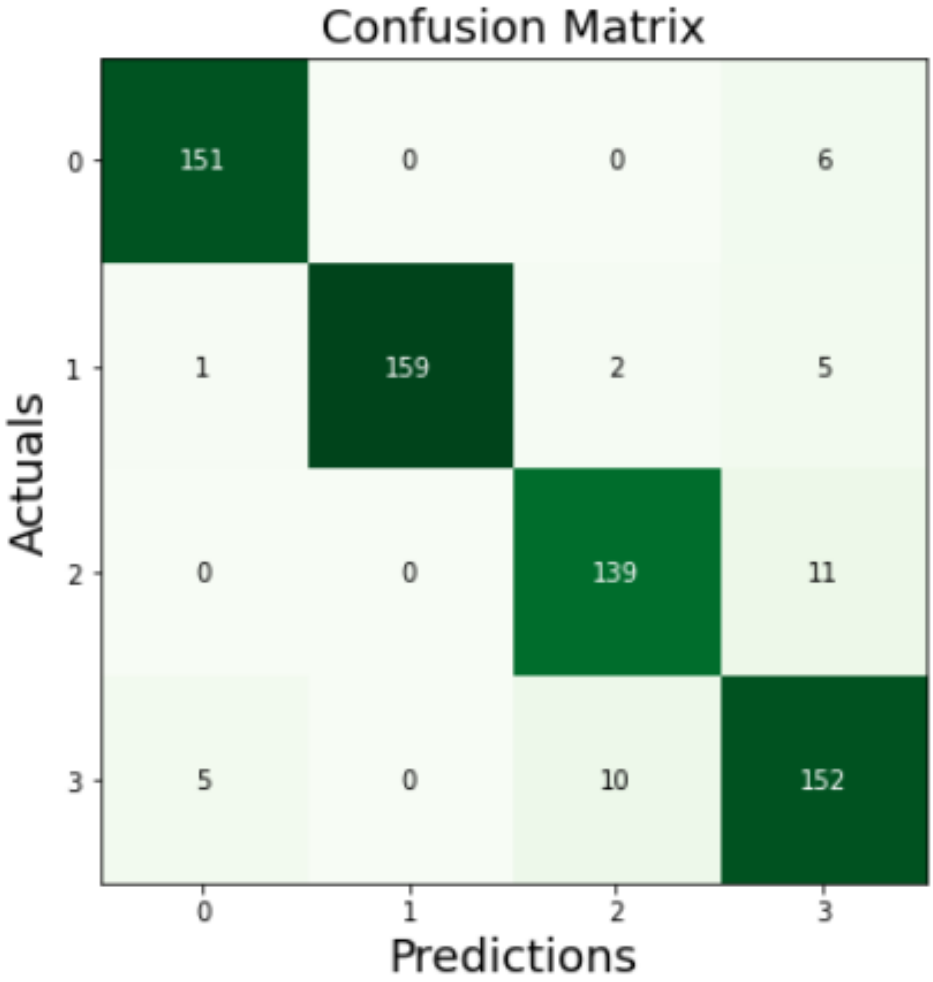

Fig. 5. Confusion matrix of MLP classifier

Table 2. Accuracies obtained by Machine Learning Algortithms

S.no	Method	Accuracy %
1	Multi layer Perceptron	93.92
2	Support vector machine	92.52
3	K Nearest Neighbors	91.35
4	Random Forest	90.71
5	Decision Trees	89.35
6	Naïve Bayes Classifier	75

The above table shows the comparison of accuracies between different machine learning models. It is seen that MLP, a neural network model is giving highest accuracy

of 93.2% followed by other machine learning models. The reason for this could be connected to fact that image processing is generally well performed by neural networks than with basic machine learning models due to the various complexities to be addressed with respect to image processing.

Table 5.2.2: Evaluation indicators of MultiLayer perceptron class prediction

Categories	Precision	Recall	F1-score
Algae water	0.96	0.95	0.95
Clean water	0.99	0.96	0.97
Mud water	0.95	0.94	0.94
Polluted water	0.87	0.92	0.89
Average	0.94	0.94	0.94

Due to the limited amount of data available, the MLP model did not yield reliable prediction results. But the proposed approach, has demonstrated a remarkable categorization effect on the featured dataset, making it appropriate for evaluating quality of water, even without a massive dataset.

6. Conclusion

Water quality plays a critical role in ensuring the efficiency of marine resources, agriculture, and human health. As a result, there is a need for fast, automated, cost-effective, and precise techniques for monitoring the quality of water. To address this need, this study proposes a water quality monitoring approach that utilizes machine learning and image processing techniques. The approach involves capturing images of water samples, including clean water, mud water, polluted water, and water with algae. The essential attributes of the water images are then extracted and used for image classification to evaluate water quality. Multiple machine learning models are employed for the image classification, and trails are carried out to determine the model with the highest accuracy. The Multi-Layer Perceptron model achieves higher accuracy of 93.92% in categorization even in the absence of extensive training samples. In the future, the model can be deployed on mobile devices to enable automatic monitoring and evaluation of quality of water across a broad range of locations. Additionally, it can serve as a primary check in water quality monitoring centers.

References

[1] Gitelson, A.A., et al. Remote estimation of chlorophyll-a concentration in inland waters using satellite red and near-infrared bands. International Journal of Remote Sensing, vol. 24, no. 18, pp. 3723-3728, 2003.

[2] Guo, H., et al. (2020). A review of big data analytics for water quality monitoring and control. Journal of Hydrology, 586, 124847.

[3] Huang, X., et al. In-situ Water Quality Monitoring Systems: A Review. Sensors, vol. 17, no. 3, pp. 613-621, 2017.

[4] Jovanović, B., et al. (2020). The use of artificial intelligence for monitoring and modeling water quality. Sustainability, 12(20), 8644.

[5] Lu, L., et al. (2020). Recent advances in the application of biosensors for environmental monitoring and early warning. Journal of Environmental Sciences, 91, 308-325.

[6] Miao, Y., et al. (2019). Deep learning for water quality assessment and prediction: A review. Science of the Total Environment, 669, 496-514.

[7] O'Neill, S. A., et al. (2017). Using drones to assess water quality in rivers and streams: A review. Science of the Total Environment, 575, 911-920.

[8] Paul, S. K., et al. (2021). An assessment of water quality indices: A review. Science of the Total Environment, 774, 145687.

[9] Satyanarayana, U. G., et al. (2020). A review on water quality sensors and their applications in urban water management. Journal of Cleaner Production, 258, 120847.

[10] T. Swapna Feature extraction to detect and classify diabetic retinopathy using fundal images International Research Journal of Engineering and Technology Vol. 9, Issue 11, pp:599-610, November 2022 Google Scholar 2395-0072

[11] United States Environmental Protection Agency. National Water Quality Monitoring Council. https://www.waterqualitydata.us/nwqmc/

[12] Xu, L., et al. (2019). A review on sensing technologies for water quality monitoring. Sensors, 19(14), 3067.

[13] Yalcin, M., et al. (2021). A review of the use of sensors and internet of things in water quality monitoring. Measurement, 169, 108398.

[14] Yuan, Y., et al. Anomaly Detection in Water Quality Data Using Machine Learning. Water Resources Management, vol. 33, no. 5, pp. 1713-1728, 2019