ECWS-8 2024 Conference

The 8th International Electronic Conference on Water Sciences

14–16 October 2024 | Online

MDPI

Water Quality Classification in terms of WQI using Machine Learning Algorithms in Keenjhar Lake, Pakistan

Muhammad Imran[1*], Danrong Zhang[1], Muhammad Zaman[2], Shazia Parveen[3], Nur E Jannat Mishu[4]

[1] College of Hydrology and Water Resources Engineering, Hohai University, Nanjing, 210098, China
[2] Department of Irrigation & Drainage, University of Agriculture, Faisalabad, 38000, Pakistan
[3] Department of Biochemistry, Bahauddin Zakariya University, Multan, 60000, Pakistan
[4] College of Information Science and Engineering, Hohai University, Changzhou, 213022, China
Correspondence author: m.imrankbr@gmail.com

INTRODUCTION & AIM

Water quality assessment is essential for monitoring the condition of water bodies. It helps to identify potential contaminants and ensure water source safety. Water Quality Index (WQI) is an effective tool to assess overall water quality. It takes into account various parameters such as temperature, pH levels, dissolved oxygen, turbidity, and levels of pollutants. These parameters are assigned specific weights and scores, which are then combined to calculate the overall WQI. Consequently, there is a pressing need to develop a comprehensive and standardized approach to calculating the water quality classification for drinking water, ensuring that it accurately reflects potential health risks and facilitates informed decision-making for both regulatory bodies and consumers.

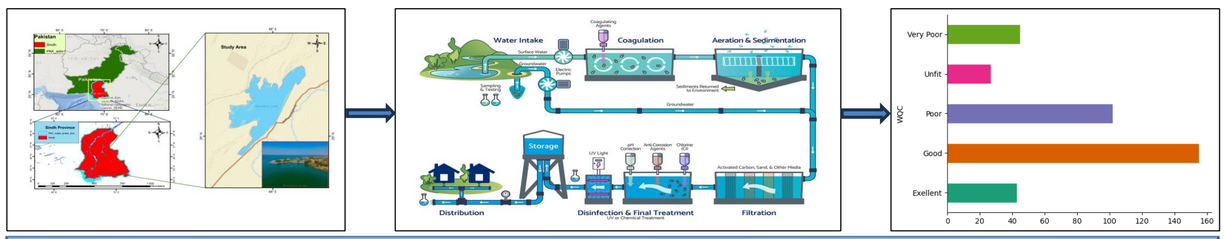
➢ **Study Area and Data Collection:-**

In this study, we selected Keenjhar Lake for our research area, it is one of the potential water supplies for drinking water to 1.8 million people of Karachi city and parts of Thatta district, Sindh, Pakistan (Lashari et al., 2014).

- Surface area ~ 13,468 ha (33,280 Acres)
- Max. L ~ 15 miles and Max. W ~ 3.7 miles
- Water Volume ~ 0.53×10^6 acre.ft (650 hm^3)

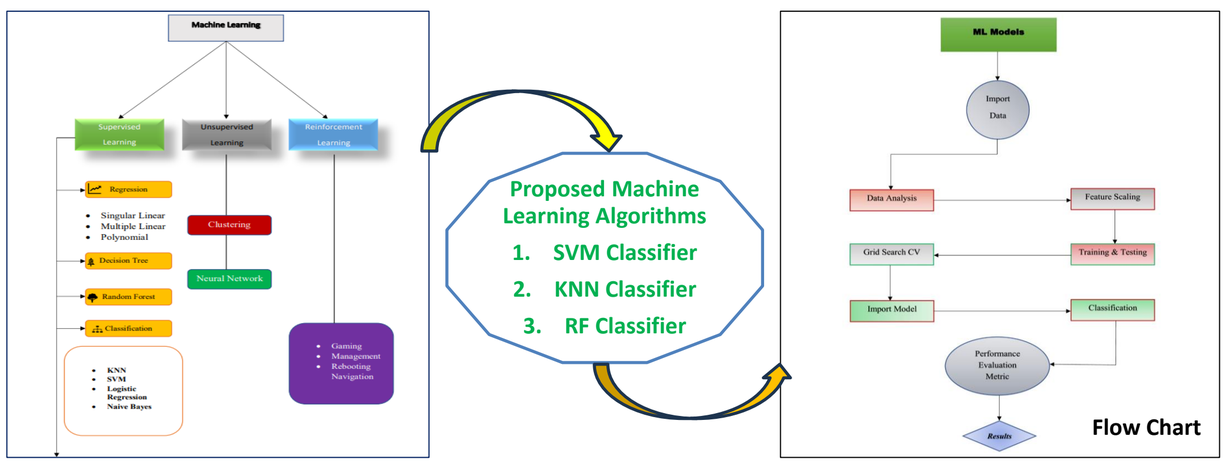
The dataset was collected from a lake at several locations in the period between 1993 to 2022.

The Irrigation Department of Sindh (Pakistan) collected this data to ensure the water is valid for drinking.



METHOD

In this study used three machine learning algorithms to predict the classification of drinking water for human beings. Machine learning is one of the most remarkable and fast-growing fields in computer science. Machine learning is built on effective algorithms that make use of a definite set of tools and functions to deal with massive and composite data sets. In this study divided the Water Quality Index (WQI) in different classes such Excellent, Good, Poor, Very Poor, and Unfit. In this study used different performance evaluation metrics such as sensitivity, specificity, f1-score, and accuracy to find the best fit algorithm.



WQI Calculations:

$$WQI = \frac{\sum_{i=1}^{N} qi*wi}{\sum_{i=1}^{N} wi}$$

$$qi = 100* \left(\frac{Vi - Videal}{Si - Videal}\right)$$

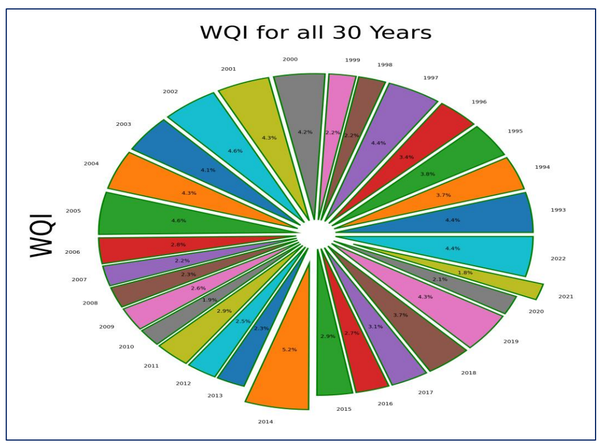
$$wi = \frac{K}{Si}$$

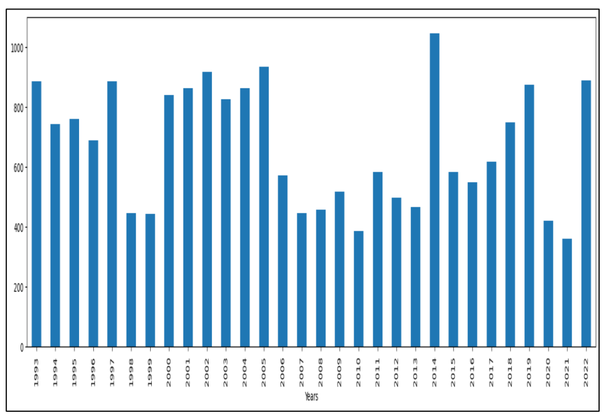
$$K = \frac{1}{\sum_{i=1}^{N} Si}$$

Maximum WQI that occurred in the year 2014 has an average value of 5.2%.

Minimum WQI in the year 2021 has an average value of 1.8%.

Parameters	Permissible Limits
Dissolved oxygen, mg/l	10
pH	8.5
Conductivity, μS/cm	1000
Biological oxygen demand, mg/l	5
Nitrate, mg/l	45
Fecal coliform, Cfu/100 ml	100
Total coliform, Cfu/100 ml	1000





RESULTS & DISCUSSION

Pearson's correlation coefficient approach is used to predict the WQI values. The correlation between the WQI parameters for selecting the optimal parameters has been obtained. Results revealed that all parameters have a strong relationship with WQI parameters. This indicates that these parameters are very important for predicting the quality of water.

$$R = \frac{n*\sum(u*v) - (\sum u)(\sum v)}{[n\sum(u^2) - \sum(u^2)]*[n\sum(v^2) - \sum(v^2)]} X100$$

where:
R: Pearson's correlation coefficient approach
u: input values in the first set of the training data
v: input values of the second set of the training data
n: total number of input variables

❖ **Splitting Dataset:**

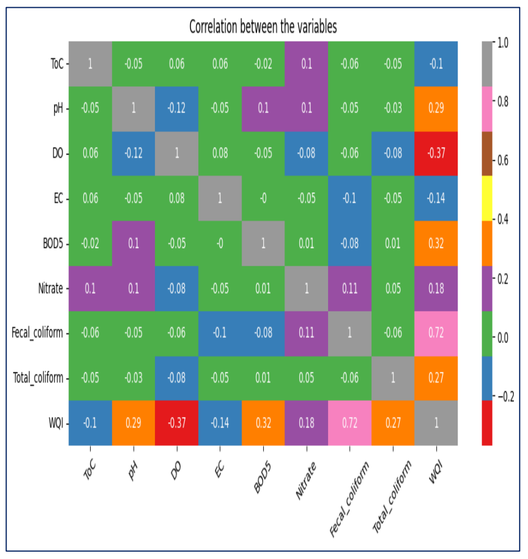
Split dataset into training and testing,
For training = 80%
For testing = 20%

$$TPR = Recall = Sensitivity = \frac{TP}{TP+FN}$$

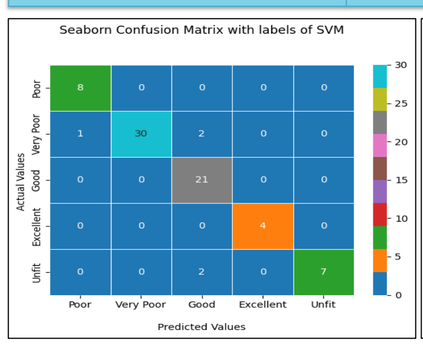
$$FPR = Precision = Specificity = \frac{FP}{FP+TN}$$

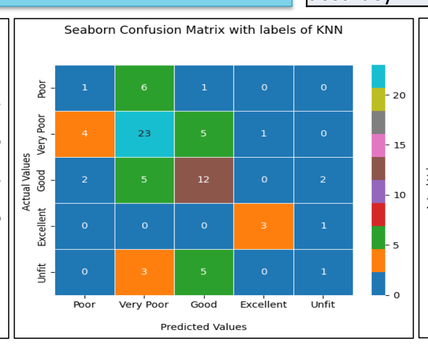
This study shows that the ROC-AUC is better for the SVM algorithm. It is also noted that the performance of the SVM algorithm is very superior as compared to the KNN and Random Forest algorithms.

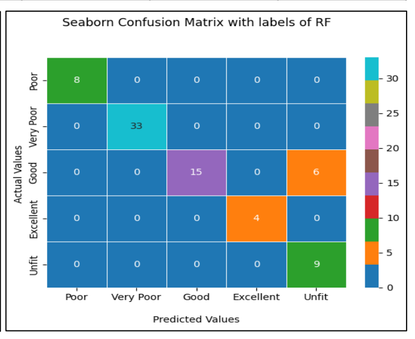
WQI range	Classification
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very Poor
>100	Unfit

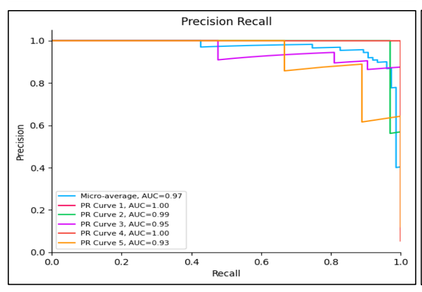


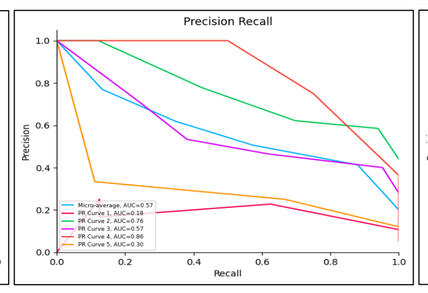
algorithms	SVM (%)	KNN (%)	RF (%)
precision	94.57	45.72	91.99
recall	93.73	45.09	94.28
f1-score	93.63	44.79	91.66
accuracy	99.5	58	95

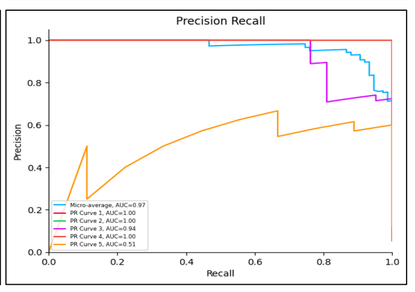


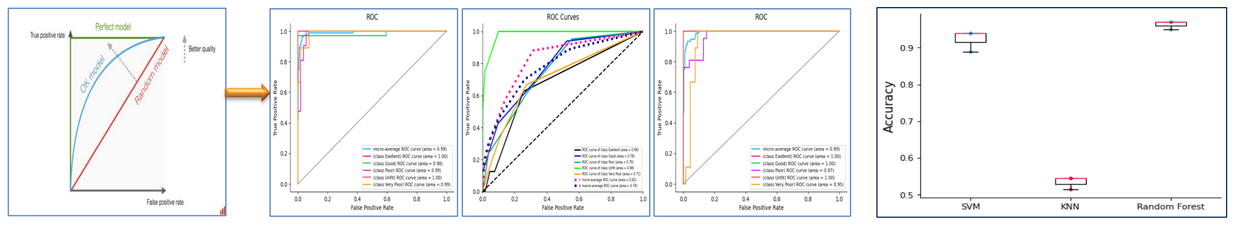












CONCLUSION

✓ In this study, the Water Quality Classification (WQC) was predicted using three different ML algorithms and the SVM algorithm was the best prediction algorithm with a higher accuracy of 99.5%.

✓ After examining the robustness and efficiency of the proposed algorithm for predicting the WQC, in future work, the developed algorithms will be implemented to predict the water quality in Pakistan for different types of water.

FUTURE WORK / REFERENCES

- WQI forecasting using a Deep Neural Network and then predicting the WQC using ML algorithms.
- Identify the most contributing region to the lake contaminations for the future.

References:

I. Lashari, K. H., Habib Naqvi, S., Palh, Z. A., Laghari, Z. A., Mastoi, A. A., Sahato, G. A., & Mastoi, G. M. (2014). The effects of physiochemical parameters on planktonic species population of Keenjhar lake, district Thatta, Sindh, Pakistan. American Journal of Bio Science, 2(2), 38–44. https://doi.org/10.11648/j.ajbio.20140202.13