ISSN: 2994-7510

Research Article

Journal of Water Research

Water Quality Classification for Precision Irrigations System Using Machine Learning and Remote Sensing

Muhammad Rashid[1]*, Shehroz Ejaz[2], Omer Farooq[1], Uzma Rafeeq[1] and Abbira Taswar[3]

[1] *MNS University of Agriculture Multan, Pakistan*

[2] *National College of Business Admission & Economics, Pakistan*

[3] *City Science & Arts College Muzaffarghr, Pakistan*

***Corresponding Author**
Muhammad Rashid, MNS University of Agriculture Multan, Pakistan.

Submitted: 2025, Jan 30; **Accepted:** 2025, Feb 03; **Published:** 2025, Feb 28

Citation: Rashid, M., Ejaz, S., Farooq, O., Rafeeq, U., Taswar, A. (2025). Water Quality Classification for Precision Irrigations System Using Machine Learning and Remote Sensing. *J Water Res, 3*(1), 01-07.

Abstract
Estimating water quality has been one of the significant challenges faced by the world in recent decades. Ensuring efficient and sustainable irrigation relies heavily on accurate water quality assessment. Contaminated water can harm soil health, crop yield, and the agricultural ecosystem. Developing precise water quality classification models is crucial, especially with the increasing demand for precision irrigation systems. This study proposes a water quality prediction model using Principal Component Regression (PCR) and Gradient Boosting Classifier (GBC). The Water Quality Index (WQI) is calculated, and Principal Component Analysis (PCA) extracts dominant water quality parameters. Several regression algorithms, including Support Vector Regression (SVR), are applied to predict WQI values. Experimental results show that the PCR model with SVR achieves 95% prediction accuracy. The Gradient Boosting Classifier achieves 100% accuracy in classifying water quality levels. The proposed approach enhances prediction reliability while reducing required parameters.

Keywords: Water Quality Classification, Irrigations System, Machine Learning, Remote Sensing

1. Introduction

Water is a fundamental resource for life and is essential for various sectors, including agriculture, industry, and domestic consumption. In irrigation systems, water quality plays a crucial role in ensuring optimal crop growth and soil health [1-3]. However, increasing industrialization and urbanization have led to severe water pollution, making water quality assessment a critical concern for sustainable agriculture. Contaminated water can introduce harmful chemicals, heavy metals, and pathogens into agricultural lands, negatively affecting crop yield and food safety [3-9].

Traditional water quality assessment methods rely on chemical and biological testing, which are often expensive, time-consuming, and require specialized laboratory facilities [11,12]. In contrast, recent advancements in machine learning (ML) offer a promising alternative by enabling automated and data-driven analysis of water quality parameters. By leveraging historical water quality data and advanced computational techniques, ML models can predict the Water Quality Index (WQI) and classify water quality status with high accuracy [13].

This study focuses on developing a machine learning-based water quality classification model for precision irrigation systems. It employs Principal Component Regression (PCR) for prediction and a Gradient Boosting Classifier (GBC) for classification to achieve high accuracy in water quality assessment. The proposed model is tested on a Gulshan Lake dataset, demonstrating its effectiveness in accurately predicting and classifying water quality. The goal is to provide an efficient, cost-effective, and scalable solution for real-time water quality monitoring in agricultural irrigation systems.

2. Literature Review

This section demonstrated the existing literature survey. The author took the most common approaches to detect and classify water quality, including deep neural networks, recurrent neural networks, neuro-fuzzy inference, and support vector regression.

Several studies have explored machine learning techniques for water quality assessment, focusing on prediction and classification models. Applied a hybrid CNN-LSTM model to predict water quality parameters such as Dissolved Oxygen (DO) and Chlorophyll-a, outperforming traditional machine learning models like Support Vector Regression (SVR) and Decision Trees [2,3,14-17]. Similarly, compared Fuzzy Logic Inference (FLI) and WQI-based models for evaluating water quality, concluding that FLI provided better accuracy [18].

For classification tasks, used a Decision Tree algorithm to classify water quality status in Klang River, Malaysia, achieving an accuracy of 84%. However, their model was limited to a small number of water quality parameters [16,17,19]. Demonstrated that Gradient Boosting and Polynomial Regression are effective for WQI prediction but noted challenges in model generalization [20,22]. Proposed a Recurrent Neural Network (RNN) combined with Dempster-Shafer Theory (DST) to analyze time-series water quality data, achieving high accuracy but requiring complex data preprocessing [1,5,19-23].

Despite these advancements, many existing models lack robustness, require extensive datasets, or focus solely on prediction rather than classification. This study addresses these gaps by combining PCA-based feature extraction, Principal Component Regression (PCR) for WQI prediction, and Gradient Boosting Classifier (GBC) for classification, ensuring improved accuracy and efficiency in water quality assessment for precision irrigation.

3. Material and Methods

This study presents a machine learning-based framework for predicting and classifying water quality in precision irrigation systems. The proposed methodology consists of four key phases: data collection, preprocessing, feature extraction, and model implementation.

3.1. Data Collection

The study utilizes a Gulshan Lake-related dataset [10], which contains multiple water quality parameters, including pH, Dissolved Oxygen (DO), Suspended Solids (SS), Electrical Conductivity (EC), Turbidity, Chloride, Chemical Oxygen Demand (COD), Total Dissolved Solids (TDS), and Alkalinity.

3.2. Data Preprocessing

To handle missing values, a median imputation technique is applied. The dataset is then normalized using a Min-Max scaler, ensuring that all variables are on a uniform scale, which improves model performance.

3.3. Feature Extraction

Principal Component Analysis (PCA) is employed to reduce dimensionality while preserving key water quality information. PCA identifies the most significant parameters influencing the Water Quality Index (WQI).

3.4. Model Implementation

Principal Component Regression (PCR) is applied using different regression models, including Support Vector Regression (SVR), Multiple Linear Regression (MLR), and Random Forest Regression (RFR), to estimate WQI values. The Gradient Boosting Classifier (GBC) is used to categorize water quality into different classes, ensuring high classification accuracy. Model performances are evaluated using metrics such as R^2 score for regression and accuracy, recall, and confusion matrix for classification.

The combination of PCA for feature reduction, PCR for WQI prediction, and GBC for classification provides an efficient and scalable approach for real-time water quality assessment in irrigation systems. Support Vector Machine (SVM) Model The SVM model was developed in 1995 by Corinna Cortes and Vapnik. It has several unique benefits in solving small samples, and nonlinear and high-dimensional pattern recognition. It can be extended to function in the simulation of other machine-learning problems. It uses the hyperplane to separate the points of the input vectors and finds the needed coefficients. The best hyperplane is the line with the largest margin, which means the distance between the hyperplane and the nearest input objects. The input points defined in the hyperplane are called support vectors. In this work, the linear SVM model along with the Gaussian radial basis function is used to classify the tested water samples based on their quality.

$$K(X,X')=exp(-||X-X'||22\sigma2)$$

where X and X' represent the feature vectors of the input dataset and the $\|X-X'\|2$ is the squared Euclidean distance between the two feature inputs. The σ is a free parameter. K-Nearest Neighbor (K-NN) Model The K-NN algorithm is a basic classification and regression method. It is used to find the K values that are close to values in the training dataset. Most of these values belong to a certain class, and thus, tested data can be classified. The K value is used to find the closest points in the feature vectors, and the value should be unique. The following expression of the Euclidean distance function (Di) can be used.

$$Di=(x1-x2)+(y1-y2)2,$$

where x1, x2, y1, and y2 are the variables for input data.

Naive Bayes Model The Bayesian method uses the knowledge of probability statistics to predict and classify datasets. The Bayesian algorithm combines prior and posterior probabilities to avoid the supervisor's bias and the overfitting phenomenon of using sample information alone.

This Naive Bayes is a type of classification algorithm based on Bayes' theorem and the assumption of the independence of characteristic conditions. Attributes are assumed to be conditionally independent of each other when the target value is given. This method greatly simplifies the complexity of the Bayesian method.

In Bayesian analysis, the probability of an event A given an event B is not the same as the probability of B given A as in equation (18).

$$P(A|B) \neq P(B|A). \quad (19)$$

Assuming that A1, A2 ⋯ . An and C are the feature vectors and the class of the WQC dataset, respectively, the Bayes equation can be expressed as follows:

$$P(C|A)=P(C)\times P(A|C)P(A)$$

where P(A) is a prior probability representing the feature vectors of the WQC dataset and P(A | C) is the prior probability of the class of the WQC dataset.

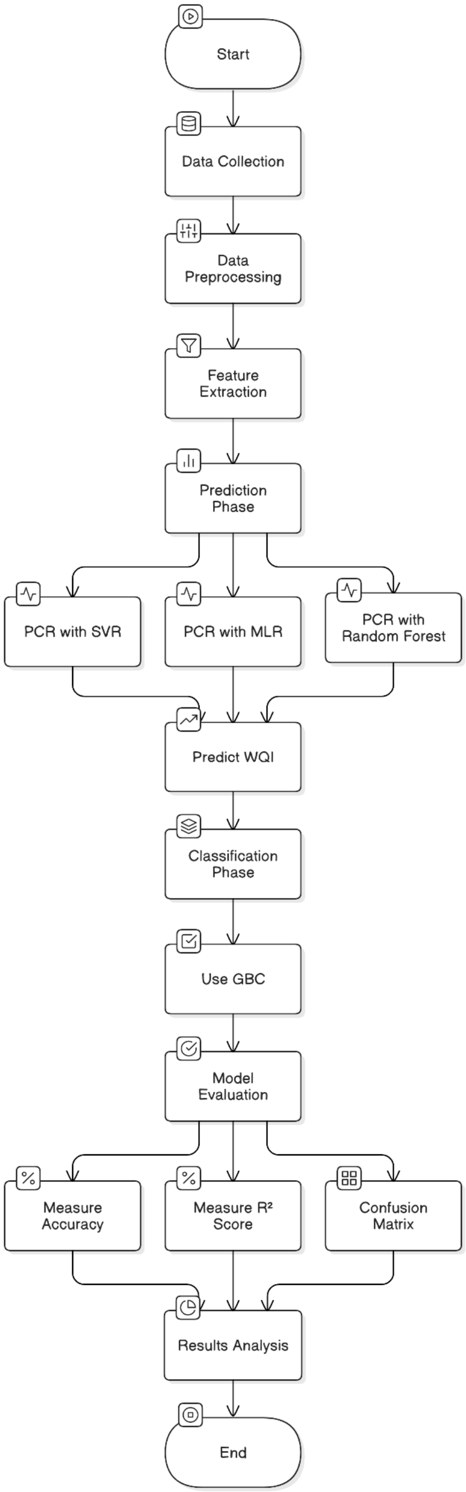

Figure 1: Overall Methology Flowchart

4. Experiments and Results

4.1. Regression Model Performance

The PCR model with Support Vector Regression (SVR) achieved the highest prediction accuracy of 95%, outperforming Random Forest Regression and Multiple Linear Regression. The PCA with the Multiple Linear Regression model also performed well, especially when the number of components was optimized.

4.2. Feature Selection Impact

A comparison of different PCA-based regression models revealed that models using eight or nine principal components (PCR8 and PCR9) achieved the best accuracy, with an R^2 score of 0.932 in the testing phase. When fewer components were used, performance dropped significantly.

4.3. Classification Model Performance

The Gradient Boosting Classifier (GBC) achieved 100% accuracy in classifying water quality levels, outperforming the Random Forest Classifier, Support Vector Classifier, and AdaBoost Classifier. The confusion matrix confirmed that GBC correctly classified all test samples, while other models showed minor misclassifications.

4.4. Comparison with Existing Models

Compared to traditional methods, the proposed approach improves both prediction and classification accuracy. While earlier studies reported classification accuracies of 71%–90%, the GBC model in this study reached 100% accuracy, demonstrating superior performance.

- PCR with SVR is the most effective model for WQI prediction, achieving 95% accuracy.
- PCA-based feature reduction helps optimize model performance while reducing computational cost.
- GBC outperforms other classifiers, achieving perfect accuracy in water quality classification.

The proposed machine learning framework provides an accurate, scalable, and cost-effective solution for water quality assessment in precision irrigation systems. Future work will focus on expanding datasets and integrating deep learning techniques to enhance model robustness.

4.5 PCR Model Result Assessment

The proposed PCR method was implemented using Python. The results of different PCR models are shown in Table 5. From this table, PCA with Support Vector Regression has achieved the highest accuracy compared to the other PCR techniques. Although other PCR models also performed well, PCA with Gradient Boosting Regression proved to be a less useful model.

Model	R^2	RMSE	MAE
PCA+ Multiple Linear Regressor	0.932	5.72	5.42
PCA+ Random Forest Regressor	0.839	8.87	7.82
PCA+ Support Vector Regressor	0.95	4.93	4.37
PCA+ Gradient Boosting Regressor	0.722	11.6	9.15

Table 1: This Table Presents the Performance Metrics (R^2, Rmse, and Mae) for Different Regression Models Combined with Principal Component Analysis (PCA).

Since the PCR model works with fewer parameters, we reduced the number of components instead of taking all the features. The results of taking different features are shown in Table 6. For this technique, PCA with Multiple linear regression is selected since PCA is mostly related to multiple linear regression to create new principal components. Table 6 illustrates that, with nine and eight components, the PCR9 and PCR8 models showed the best performance, where PCR9 clarified all the variance. The PCR8 model gives the same result as the PCR9 model, and the number of parameters is also reduced.

The R^2 value for the PCR8 model in the testing steps is .932. If we reduce one more component from the PCR8 model, that model produces almost the same result as operating with all the components. The R^2 value in the PCR7 model is .927. After reducing one more component, the R^2 value reduced in the testing phases is .709. That shows less accuracy compared with the PCR7 and PCR8 models. Yet in the water samples, PCR6 still performed well. If we reduce more components from the PCR model, the R^2 value is barely 50 percent, which shows low PCR model efficiency.

Model	Components	R^2	RMSE	MAE
PCR1	n=1	0.042	21.64	19.56
PCR2	n=2	0.411	16.97	14.57
PCR3	n=3	0.589	14.52	12.51
PCR4	n=4	0.564	14.59	12.54
PCR5	n=5	0.565	14.58	11.67
PCR6	n=6	0.709	11.92	9.44
PCR7	n=7	0.927	5.97	5.33
PCR8	n=8	0.932	5.72	5.42

PCR9	n=9	0.932	5.72	5.42

Table 2: This Table Presents the Performance Metrics (R², Rmse, and Mae) for Principal Component Regression (Pcr) Models Using Different Numbers of Components.

The accuracy comparison of the PCR model in each principal component is shown in Figure 3. The model performed well with six, seven, and eight components. After that, it showed poor performance. Since PCR7 and PCR8 showed the same results as working with PCR9, we could infer that the PCR method allows operating with fewer parameters instead of taking all the features.

Figure 2, illustrates the plot between the observed and predicted WQI values for a better understanding of those models. Among them, the value appeared closer to the regression fit line in the PCA+ Support Vector Regression model because of the high training and testing accuracy.

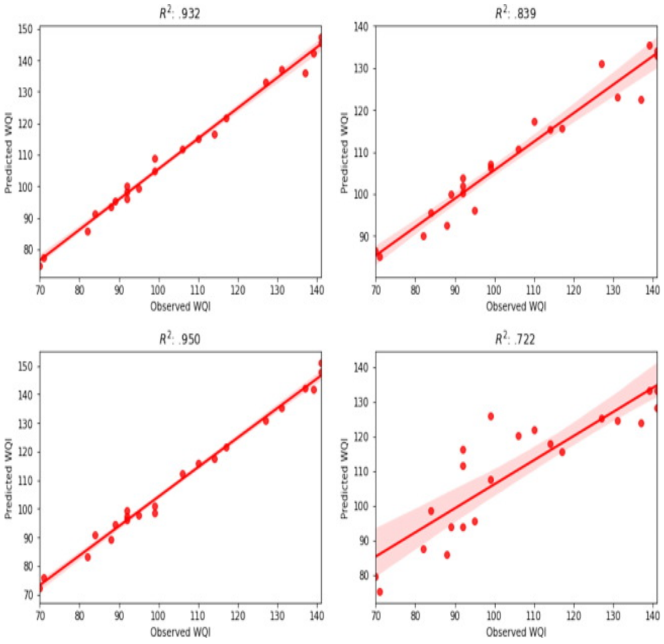

Figure 2: Prediction and Classification of Water Quality Indexplot Between Observed and Predicted WQI a. PCA+ Multiple Linear Regression Model b. PCA+ Random Forest Regression Model c. PCA+ Support Vector Regression Model d. PCA+ Gradient Boosting Regression Model.

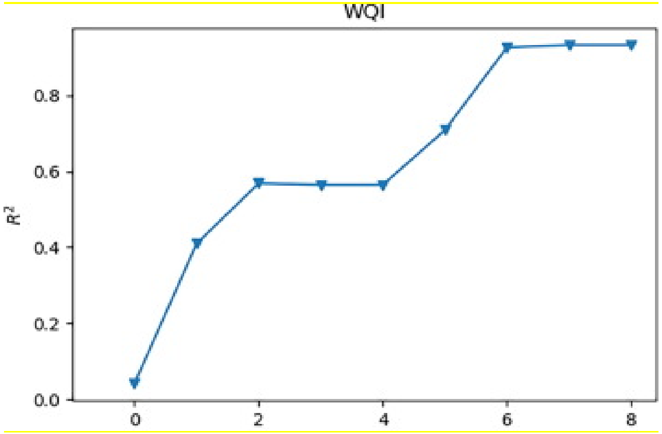

Figure 3: The Number of Principal Components is the Regression Model

4.5.Classification Model Result Assessment

Different classification algorithms are implemented using Python. The results of varying classification models are presented in 3. Among them, the Gradient Boosting Classifier has achieved the highest accuracy and proved to be an efficient model to predict water quality status. The second-best model is the Random Forest Classifier, but to calculate recall, the Support Vector Classifier performs better than the Random Forest Classifier. Ada-Boost Classifier is found less effective model compared to the other techniques. In Figure 4, the confusion matrix for those models is presented we can observe that the Gradient Boosting Classifier classifies all the testing data according to the water quality level whereas other models misclassifies some of the testing data.

Model	Accuracy	Precision	Recall	F1-score
Random Forest Classifier	0.91	0.96	0.85	0.89
Support Vector Classifier	0.86	0.82	0.91	0.84
Gradient Boosting Classifier	1	1	1	1
Adaboost Classifier	0.77	0.53	0.6	0.56

Table 3: This Table Compares the Performance Metrics (Accuracy, Precision, Recall, and F1-Score) of Different Classification Models.

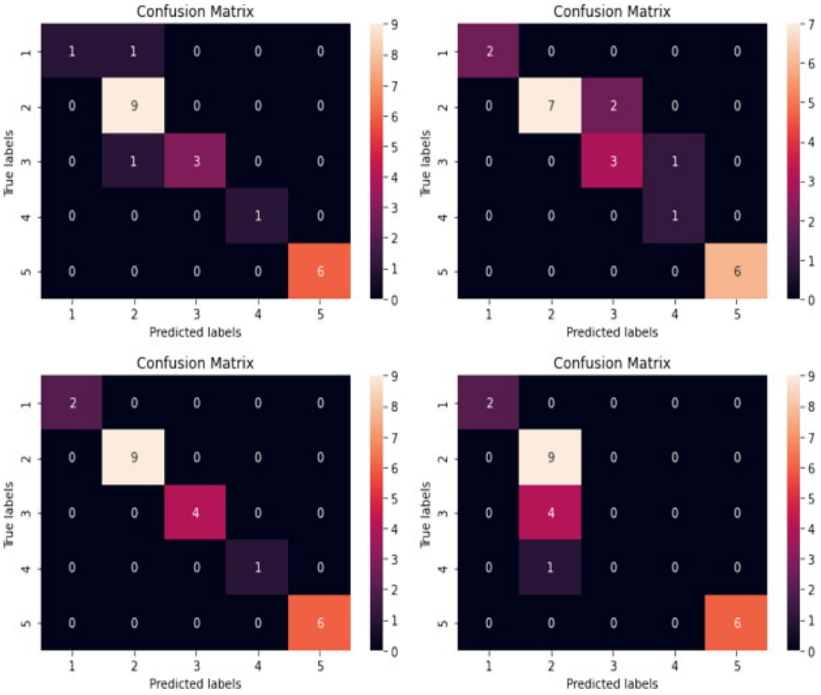

Figure 4: Confusion Matrix of Overall Models

5. Discussion

The results of this study highlight the effectiveness of machine learning models in predicting and classifying water quality for precision irrigation systems. The Principal Component Regression (PCR) with Support Vector Regression (SVR) demonstrated 95% accuracy, making it a reliable approach for Water Quality Index (WQI) prediction. Compared to traditional regression models, PCR significantly reduced dimensionality while maintaining high predictive accuracy. This indicates that feature selection through PCA plays a crucial role in improving model efficiency.

For classification, the Gradient Boosting Classifier (GBC) achieved 100% accuracy, outperforming other models such as Random Forest and Support Vector Classifier. The confusion matrix analysis confirmed that GBC correctly classified all test samples, indicating its strong generalization capability. The high classification accuracy suggests that boosting techniques are highly effective in handling complex water quality datasets. These findings align with previous research, where deep learning and ensemble learning models demonstrated superior performance in water quality assessment. However, this study achieves high accuracy with fewer computational resources, making it suitable for real-time applications in precision irrigation. Future research should explore deep learning-based models and integrate IoT sensors to enable continuous water quality monitoring in agricultural settings.

6. Conclusion

This study presents an efficient machine learning-based framework for water quality prediction and classification in precision irrigation systems. The proposed model combines Principal Component Regression (PCR) for WQI prediction and Gradient Boosting Classifier (GBC) for classification, achieving 95% and 100% accuracy, respectively. Key findings indicate that PCA-based feature extraction optimizes model performance, allowing for accurate predictions with fewer computational requirements. The GBC model outperformed other classifiers, confirming its robustness in classifying water quality levels. Compared to existing studies, this approach provides higher accuracy, reduced dimensionality, and better generalization, making it ideal for real-time water quality assessment. The results demonstrate that machine learning techniques can effectively address water quality challenges in irrigation systems, helping farmers make data-driven decisions for sustainable agriculture.

Future research should focus on expanding datasets, incorporating deep learning models, and integrating IoT-based real-time monitoring to further enhance prediction accuracy and scalability. By adopting such intelligent systems, agriculture can move towards smarter, more sustainable water management solutions.

References

1. Kukartsev, V., Orlov, V., Semenova, E., & Rozhkova, A. (2024). Optimizing water quality classification using random forest and machine learning. In *BIO Web of Conferences* (Vol. 130, p. 03007). EDP Sciences.
2. Vavekanand, R., Sathio, A. A., Singh, V., & Anwar, S. (2024). Water4. 0: An Industrial Water Pollution Forecasting Using Machine Learning. *Available at SSRN 4849924.*
3. Yao, Z., Wang, Z., Huang, J., Xu, N., Cui, X., & Wu, T. (2024). Interpretable prediction, classification and regulation of water quality: A case study of Poyang Lake, China. *Science of the Total Environment, 951*, 175407.
4. Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications, 83*(12), 35307-35334.
5. Jayaraman, P., Nagarajan, K. K., Partheeban, P., & Krishnamurthy, V. (2024). Critical review on water quality analysis using IoT and machine learning models. *International journal of information management data insights, 4*(1), 100210.
6. Vavekanand, R., & Kumar, S. (2024). Rural Agricultural Development Through E-Commerce Platforms. *Authorea Preprints*.
7. Ikhlef, N., Tachi, S. E., Bouguerra, H., Djabri, L., & Arrar, J. (2024). Classification of Groundwater Quality for Irrigation Purposes in Wetland Region by Irrigation Water Quality Index. *Water Resources, 51*(3), 322-331.
8. Vavekanand, R. (2024). A Machine Learning Approach for Imputing ECG Missing Healthcare Data. *Available at SSRN 4822530.*
9. Pérez-Beltrán, C. H., Robles, A. D., Rodriguez, N. A., Ortega-Gavilán, F., & Jiménez-Carvelo, A. M. (2024). Artificial intelligence and water quality: From drinking water to wastewater. *TrAC Trends in Analytical Chemistry*, 117597.
10. Mohinuzzaman, M., Kamrujjaman, M., Hossain, S. M., & Saadat, A. H. M. (2013). Quality assessment of water and sediment of Gulshan lake by using neutron activation analysis. *Jahanginagar University. Bangladesh*. pp-12-15.
11. Arepalli, P. G., & Naik, K. J. (2024). Water contamination analysis in IoT enabled aquaculture using deep learning based AODEGRU. *Ecological Informatics, 79*, 102405.
12. Tian, D., Zhao, X., Gao, L., Liang, Z., Yang, Z., Zhang, P., ... & Chen, J. (2024). Estimation of water quality variables based on machine learning model and cluster analysis-based empirical model using multi-source remote sensing data in inland reservoirs, South China. *Environmental Pollution, 342*, 123104.
13. Vavekanand, R. (2024). Advancements in Software Engineering for IoT Applications: Addressing Challenges and Seizing Opportunities.
14. Vavekanand, R. (2024). From Language to Action: A Study on the Evolution of Large Language Models to Large Action Models. *Authorea Preprints*.
15. Vavekanand, R., & Dayanand, R. (2024). Digital Agri: Bridging the Gap for Equitable Access to Technology in Rural Communities. *Available at SSRN 4810580*.
16. Mondal, I., Hossain, S. A., Roy, S. K., Karmakar, J., Jose, F., De, T. K., ... & Nguyen, N. M. (2024). Assessing intra and interannual variability of water quality in the Sundarban mangrove dominated estuarine ecosystem using remote sensing and hybrid machine learning models. *Journal of Cleaner Production, 442*, 140889.
17. Jayaraman, P., Nagarajan, K. K., Partheeban, P., & Krishnamurthy, V. (2024). Critical review on water quality analysis using IoT and machine learning models. *International journal of information management data insights, 4*(1), 100210.
18. Wei, H., Jia, K., Wang, Q., Cao, B., Qi, J., Zhao, W., & Yan, K. (2024). A remote sensing index for the detection of multi-type water quality anomalies in complex geographical environments. *International Journal of Digital Earth, 17*(1), 2313695.
19. Vavekanand, R., Sam, K., Kumar, S., & Kumar, T. (2024). Cardiacnet: A neural networks based heartbeat classifications using ecg signals. *Studies in Medical and Health Sciences, 1*(2), 1-17.
20. Ma, Y., Chen, S., Ermon, S., & Lobell, D. B. (2024). Transfer learning in environmental remote sensing. *Remote Sensing of Environment, 301*, 113924.
21. Aslam, R. W., Shu, H., Javid, K., Pervaiz, S., Mustafa, F., Raza, D., ... & Hatamleh, W. A. (2024). Wetland identification through remote sensing: insights into wetness, greenness, turbidity, temperature, and changing landscapes. *Big Data Research, 35*, 100416.
22. Zhu, L., Cui, T., Runa, A., Pan, X., Zhao, W., Xiang, J., & Cao, M. (2024). Robust remote sensing retrieval of key eutrophication indicators in coastal waters based on explainable machine learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 211*, 262-280.
23. Vavekanand, R., Sam, K., & Singh, V. (2024). UAV Networks Surveillance Implementing an Effective Load-Aware Multipath Routing Protocol (ELAMRP). *arXiv preprint arXiv:2407.09531.*

Copyright: *©2025 Muhammad Rashid, et al . This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*