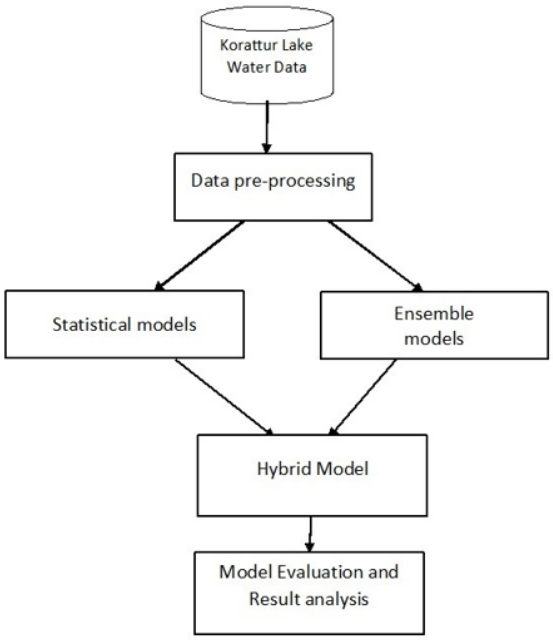
Global NEST Journal, Vol 26, No 2, 05492

Water quality prediction using statistical, ensemble and hybrid models

D. Venkata Vara Prasad, Lokeswari Y Venkataramana, Vikram V, Vyshali S. and Shriya B
Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, Chennai, India
Received: 05/11/2023, Accepted: 04/02/2024, Available online: 13/02/2024
*to whom all correspondence should be addressed: e-mail:
https://doi.org/10.30955/gnj.005492

Graphical abstract



Abstract

Water is an essential elixir for several living organisms to function and survive. But it gets contaminated through several sources such as industrial wastes, oil spills, marine dumping, etc. With a growing population, availability of good quality water is of grave importance. This has become the motivation to probe into analysis of water quality from the outcomes of Statistical and Ensemble methods and to find the best working models from both methods. Research has been done to predict water quality analysis using standalone statistical and ensemble models. So, this research focuses on obtaining the best Statistical and Ensemble model separately among the models tried. The statistical models implemented for comparison are Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA). The Ensemble models used are Bagging, Boosting and Stacking. The models are then combined to build a Hybrid model to observe the comparisons between the three. The performance metrics used are Confusion Matrix, Accuracy, Precision, Recall, F1-score and ROC curve. While comparing the models, it is observed that Hybrid model produces the most accurate results, hence proving that the combination of Statistical and Ensemble model is efficient.

1. Introduction

Water bodies have played a critical role in personal and industrial uses. Polluted water bodies can lead to series of infection and sometimes could even lead to death. Several industries like fabrication, food, agriculture, automotive and so on, require water quality prediction tools with great prediction ability in order to manufacture products with good quality. This research aims to analyse the water quality to make sure that good quality water is available to drink and use for all purposes. Traditional methods like lab analysis have several shortcomings and could consume a lot of time. This research adopts innovative techniques to solve such shortcomings. This research helps to get an insight on how combining two best working models can produce better results than standalone models.

Statistical models are mathematical techniques and statistical assumptions that generate sample data and make predictions. It usually is a collection of probability distributions on a set of all possible outcomes of an experiment. The Statistical models used in this research are Principal Component Analysis, Hierarchical Clustering Analysis, Quadratic Discriminant Analysis and Linear Discriminant Analysis. Ensemble methods create multiple models and combine them to produce better results.They usually produce solutions that are higher in accuracy than a single model. The Ensemble models implemented are Bagging, Boosting and Stacking. The statistical and ensemble learning models are combined to form a hybrid model. The outcome of the hybrid model is compared with ensemble learning and statistics based systems in order to analyse the performance of the hybrid model.

Pham *et al.* (2020) implemented data intelligence models along with ensemble methods in order to predict water quality index. The algorithms used were BPNN, ANFIS, SVR and MLR to predict the WQI of three stations- Nizamuddin, Palla, and Udi along the river Yamuna. The results indicated that NNE was the best approach for the prediction of water quality index.

Nguyen Thi Thuy Linh [Sani Isah *et al.* 2020] proposed a system that applied AI based models like LSTM, ELM, GRNN and HW along with ensemble models such SAE and WAE (linear) and BPNN-E and HW-E (non-linear) and a hybrid random forest ensemble for prediction of dissolved oxygen in water. All the hybrid models showed great results but HW-RF ensemble seemed to be the best.

Rodelyn *et al.* (2018) performance of statistical model in water quality prediction. The statistical models used were naive model, multiple linear regression, dynamic regression, regressiontree, Markov chain, classification tree, random forests, multinomial logistic regression, discriminant analysis and Bayesian network. Results indicated that Bayesian network had the best performance.

Rahim Barzegar *et al.* (2018) researched on multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. Results indicated that the hybrid version had greater performance than the individual models. Between the two hybrids, WA-ELM had better performance.

Xingguo Chen *et al.* (2021) identified suitable model for water quality prediction among traditional, ensemble, cost-sensitive, outlier detection learning models and sampling algorithms. The Traditional models used were decision tree, Logistic regression, K-nearest neighbour, and support vector machine. The ensemble techniques chosen were RF, DCF, and gradient boosting decision tree. DCF was found to be best performing among all the chosen models. Cost sensitive RF and AdaCost were also observed to produce excellent results.

Gozen Elkiran *et al.* (2019) researched on Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. The AI models used were BPNN, ANFIS, SVM and a traditional linear model ARIMA along with three other ensemble techniques to enhance their performances. The performance metrics used were Determination coefficient and root mean square error. For SL1, ANFIS has better results, for SL2, ANFIS had better performance too. However for SL3, SVM had better results compared to others. To increase performance, ensembling was used. It was recorded that NNE was proved the most effective of all other techniques.

Zengrui *et al.* (2019) attempted to design a data pre-processing model based on statistical detection methods. The statistical methods chosen were quartile detection method and Z-score method. The main functionality was based on quartile detection method, but its screening result was corrected by the Z-score method. The result after screening was processed into a matrix format.

Y. Khan *et al.* (2017) proposed an ensemble of ANN and ANFIS for water quality prediction and analysis. The research used a hybrid of ANN and ANFIS in order to record the prediction accuracy. Results indicated that the ANN-ANFIS model was the most accurate with greater prediction accuracies.

Ozgur Kisi *et al.* (2020) proposed a new ensemble model called Bayesian model averaging to predict dissolved oxygen lecvels in water. The proposed model was compared with extreme learning machine, artificial neural network, adaptive neuro-fuzzy inference system, classification and regression tree and multilinear regression. The proposed model performed best.

Li-ming (Lee) he *et al.* (2008) developed an ANN model to predict TC, FC and EN. The model was found to have great performance with fast prediction ability. They also indicated that the model is ready to be employed to other coastal beaches too.

A. Najah *et al.* (2011) used models such as multi-layer perceptron neural networks, ensemble neural networks and support vector machine. It was cited that SVM was found to overcome all the drawbacks of the other models.

Navideh Noori *et al.* (2020) researched on water quality prediction using SWAT-ANN coupled approach. The research indicated that hybrid models can be more robust and accurate than standalone models.

Sanghyun Park *et al.* (2020) researched on variable update strategy to improve water quality

forecast accuracy in multivariate data assimilation using ensemble kalman filter. The Hydrologic Simulation Program-Fortran (HSPF) model and the Environmental Fluid Dynamics Code (EFDC) model were employed. The case which had CHL, PO4 3–P, NH4+-N, and DO as parameters was found to be the best performing case.

Ali Omran Al-sulthani *et al.* (2021) proposed ensemble data-intelligence models for surface water quality prediction. The ensemble models used were Quantile regression forest, random forest, radial support vector machine, stochastic gradient boosting and gradient boosting machines. The results indicated that the QRF model seemed to have the maximum performance.

Leizhi Wand *et al.* (2021) researched on improving robustness of Beach water quality modeling using Stacking, an ensemble approach. The outcome of five common individual machine learning models - multiple linear regression, partial least square, sparse partial least square, random forest, and Bayesian network were taken as input for another model that gave the final prediction. The performance metrics used were Cross validation, MSE and accuracy. It was observed that different models performed well for different samples.

Abobakr Saeed Abobakr Yahya *et al.* (2019) developed a model using SVM to predict water quality. The model was found to be very efficient and robust.

D. Venkata Vara Prasad *et al.* (2020) explored different types of machine learning algorithms to predict the water quality index and the water quality class. The ML models used were support vector machine, decision tree, logistic regression, random forest, and naive Bayesian. The performance metrics were accuracy and precision. Among all the algorithms, the random forest algorithm produced an accuracy of 95% which was the highest with least execution time.

D. Venkata Vara Prasad *et al.* (2021) researched on Machine Learning algorithms for comparing AutoML and an expert

architecture built by the authors to evaluate the Water Quality Index and the Water Quality Class. The results indicated that the accuracy of AutoML and TPOT was 1.4% higher than conventional ML techniques for binary class water data. In the case of Multi class water data, AutoML was 0.5% higher and TPOT was 0.6% higher than conventional ML techniques. (2021)

Venkata Vara Prasad D *et al.* (2021) explored many deep learning algorithms to predict the Water Quality Index and the Water Quality Class. The dataset was collected from Korattur Lake in the Chennai city, Tamilnadu. . The deep learning models used were Artificial Neural Network, Recurrent Neural Network and Long-Short Term Memory. The performance metrics used were accuracy, precision and the execution time. The results indicated that LSTM produced the highest accuracy of 94% and also consumed the least execution time. (2020)

Shuangyin Liu *et al.* (2012) presented a hybrid model called real-value genetic algorithm support vector regression (RGA–SVR), which searched for the optimal SVR parameters using real-value genetic algorithms, to further construct the SVR models. The results showed that

RGA–SVR performed better than the traditional SVR and back-propagation (BP) neural network models based on the root mean square error (RMSE) and mean absolute percentage error (MAPE).

Yunrong Xiang *et al.* (2009) dealt with water quality prediction model through application of LS-SVM. LS-SVM along with particle swarm optimization (PSO) was used for time series prediction. Testing the model showed high efficiency in predicting the water quality of the Liuxi River.

Chenguang Song *et al.* (2022) proposed a hybrid model based on the ensemble learning method that combined the entire ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and improved LSTM to predict the water quality parameters. The results showed that the proposed model had greater accuracy than models.

Zilin Li *et al.* (2022) presented a new stacking ensemble model for detection of water quality using multiple parameters. The stacking method had a higher true positive rate, lower false positive rate and higher F1 score.

Park Jungsu (2021) developed an ensemble machine learning model to predict Suspended sediment concentration using the XGBoost (XGB) algorithm. The RSR were 0.51 and 0.57 in the two monitoring stations for Model 2, respectively, while the model performance improved to RSR 0.46 and 0.55, respectively, for Model 1.

Park Jungsu (2022) developed an XGBoost ensemble machine learning (ML) model from 18 input variables to predict Chl-a concentration.This study successfully demonstrated a good example of XAI application to improve the ML model performance in predicting water quality. Lingbo Li *et al.* (2022) evaluated five tree-based models- classification tree, random forest, CatBoost, XGBoost, and LightGBM, and employed an explanation method SHAP to explain the models used. The results suggested that the combination of LightGBM and SHAP had good potential to develop interpretable models for predicting microbial water quality in freshwater lakes.

Farid Hassanbaki Garabaghi *et al.* (2021) analysed the performance of four machine learning algorithms with ensemble learning approach and proposed a classifier with highest performance. Three feature selection methods employing machine learning were applied.As a result XGBoost classifier was suggested as the best classifier with the maximum accuracy of 95.606%.

Arshia Fathima *et al.* (2014) the present study focuses on devising a prediction model for BOD using ensemble techniques in data mining. A correlation coefficient of 0.9541 was obtained for the proposed model. Bagging for the river data showed good results without over-fitting.

Rosaida Rosly *et al.* (2022) researched methods boosting, bagging, and stacking. The result showed that the stacking method with MLP algorithm achieved higher accuracy of 96.39%.

Dipankar Ruidas *et al.* (2022) researched on Bagging, random forest (RF), and an ensemble of bagging and RF were employed to assess the HHRM. Performance was analysed using several statistical validating methods. The results showed that ensemble technique was best performing.

The authors Lateko, A. A *et al.* (2021) proposed an approach for one-day to three-day ahead PV power hourly forecasting based on the stacking ensemble model with a recurrent neural network (RNN) as a meta-learner.

Li, Z *et al.* (2022) The stacked ensemble model was constructed with machine learning base models and meta-learners trained using cross validation to identify important water quality parameters and detect the water contamination in water distribution system.

Reddy, P *et al.* (2023) The quality of Continuous Assessment Tests (CAT) question papers was better understood using machine learning techniques.

Dietterich, T. G *et al.* (2000) compares the effectiveness of randomization, bagging, and boosting for improving the performance of the decision-tree algorithm C4.5. Randomization is slightly better than bagging but not accurate as boosting.

Mosavi, A *et al.* (2021) compared the four ensemble models, Boosted generalized additive model (GamBoost), adaptive Boosting classification trees (AdaBoost), Bagged classification and regression trees (Bagged CART), and Random Forest (RF). Ground water quality was predicted using these ensemble models. Recursive Feature Elimination (RFE) was used to identify prominent ground water parameters. Bagging models (i.e., RF and Bagged CART) had a higher performance than the Boosting models (i.e., AdaBoost and GamBoost). Random Forest out performed other models with 86% accuracy.

2. System design

2.1. Dataset collection

Out of 4 datasets under study, three are taken from Korattur Lake, located in Chennai, capital of Tamil Nadu, one of the states in South India. The Korattur Lake occupies an area of 990 acres and has always been one of the major sources of drinking water. The dataset has observations made over a 12-year period from 2009 to 2021. The three datasets from Kaggle have 9 attributes- pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD (mg/L), Chlorine and Sodium. The last dataset is obtained from Kaggle, an open-source environment, which is a subsidary of Google. Kaggle allows users to upload as well as download datasets as well as publish models in a data science environment which is totally web based.

2.2. Datasets description

2.2.1. Binary Class Dataset from Korattur Lake

The rows in the dataset are classified into two classes 0 and 1 based on 9 attributes. The size of the dataset is 5001 × 10, that is it has 5001 rows and 10 columns including the class label.

2.2.2. Binary class dataset from kaggle

The rows in the dataset are classified into two classes 0 and 1 based on 20 attributes. The 20 attributes are - Aluminium, Ammonia, Arsenic, Barium, Cadmium, Cloramine, Chromium, Copper, Fluoride, Bacteria, Lead, Nitrates, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver and Uranium. The size of the dataset is 8000 × 21, that is it has 8000 rows and 21 columns including the class label.

2.2.3. Three class dataset from korattur lake

The dataset has three classes namely 0, 1 and 2 where 0 indicates that the quality of water is excellent, 1 indicates that the water is good, and 2 indicates that the water quality is poor. The size of the dataset is 10140 × 10, that is it has 10140 rows and 10 columns including the class label.

2.2.4. Five class dataset from korattur lake

The dataset consists of five classes where 0 stands for excellent, 1 stands for good, 2 stands for average, 3 stands for bad and 4 for poor water quality. The size of the dataset is 5100 × 10, that is it has 5100 rows and 10 columns including the class label.

2.3. Data splitting

To find out the performance of the Models to be implemented, it is important to split the data into training and testing data. After dataset splitting, the model is trained with the training data and then tested on parts of data for finding accuracy of the model for its performance analysis. The dataset was split in the ratio of 4:1 for training and testing respectively.

2.4. Exploratory data analysis

2.4.1. Binary Class Dataset from Korattur Lake

According to the box plot made on the features of the Binary Class dataset sourced from Korattur Lake as shown in Figure 1, it is derived that there are no outliers on any of the features.

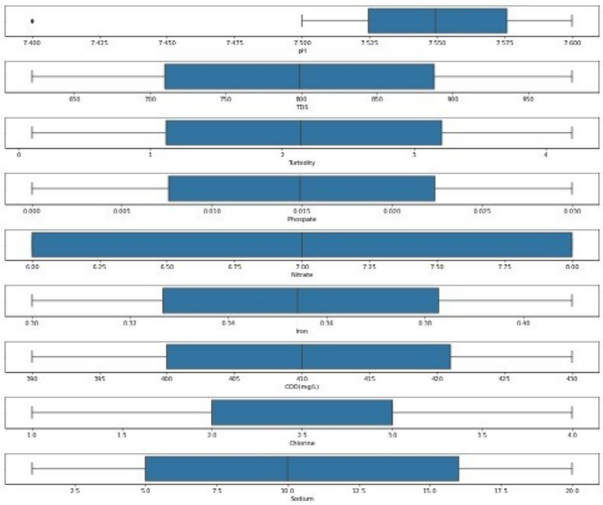


Figure 1. Box plot of Binary Korattur dataset

2.4.2. Binary class dataset from kaggle

According to the radar plots made on the features of the Binary Class dataset sourced from Kaggle as shown in Figures 2 and 3, it is derived that there are some outliers on all the features including Aluminium, Ammonia, Arsenic, Barium, Cadmium, Cloramine, Chromium, Copper, Fluoride, Bacteria, Lead, Nitrates, Nitrites, Mercury, Perchlorate, Radium, Selenium, Silver and Uranium.

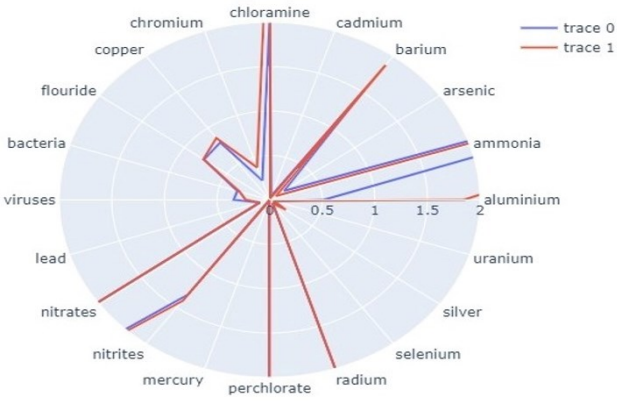


Figure 2. Binary Kaggle data-set - Overview from range 0 to 2

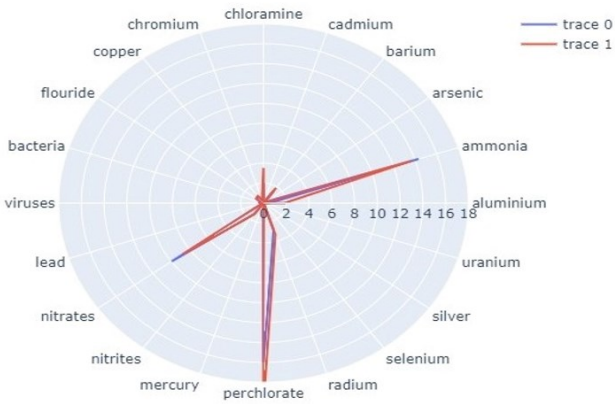


Figure 3. Binary Kaggle data-set - Overview from range 0 to 18

2.4.3. Three Class Dataset from Korattur Lake

According to the box plot made on the features of the three Class dataset sourced from Korattur Lake as shown in Figure 4, it is derived that there are no outliers on any of

the features including pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD (mg/L), Chlorine and Sodium.

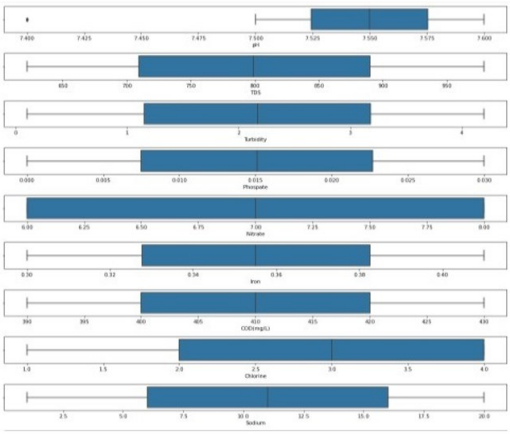


Figure 4. Boxplot for Three Class Korattur Lake dataset

2.4.4. Five Class Dataset from Korattur Lake

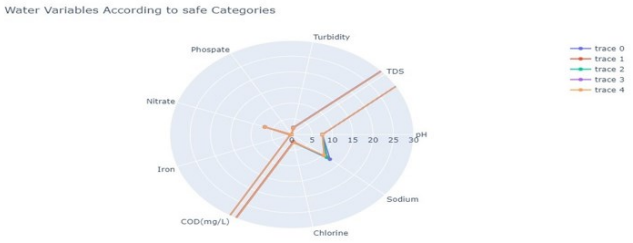


Figure 5. Five class Korattur Lake dataset

According to the radar plot made on the features of the five class dataset sourced from Korattur Lake as shown in Figure 5, it is derived that there are no outliers on any of the features including pH, TDS, Turbidity, Phosphate, Nitrate, Iron, COD (mg/L), Chlorine and Sodium.

2.5. Data cleaning

There were no missing values in any of the datasets, so missing values were not taken care of. The data did not have any outliers, so the data was noise-free. The data had no structural errors, unwanted outliers, missing data so there was no major cleaning to be done.

2.6. Models comparison and analysis

The Statistical and Ensemble models were compared in parallel and the results were analysed to obtain the most efficient Statistical and Ensemble model separately. The conclusions were drawn based on the data in hand in context of the case being studied.

2.7. Hybrid model

The architecture diagram for the hybrid model is shown in Figure 6. The flow of the data is seen to first go to the selected Statistical model, then to the Ensemble model. The data is trained and tested through these two models which combine to form the hybrid model. The Hybrid model is then used to predict the result.

3. Implementation

This research engages two parallel techniques, such as Statistical models and Ensemble models, and applies to the data. Various Machine and Deep Learning techniques are combined to construct the Ensemble models. They are built, trained, and the results are obtained. Using the results, conclusions are drawn to select the best performing Statistical and Ensemble model to build the Hybrid Model.

3.1. Statistical techniques

3.1.1. Principal component analysis

PCA is used for dimensionality reduction, (i.e) reduces the feature space by removing noisy and unclean features from a real world dataset. Thus makes the dataset easier to visualize, analyze and interpret. Two classes come under PCA - Feature Elimination, Feature Extraction. In this research, feature extraction is used.

3.1.2. Hierarchical clustering analysis

HCA's objective is to group several features/data points in such a way that they are close to one another. The fundamental technique is to repeatedly calculate the distance between the features and further calculate the distances between the clusters once the features/data points start forming clusters. The outputs are usually represented as a dendrogram. Two methods that fall under HCA are Divisive methods and Agglomerative methods. Here, Agglomerative method is used.

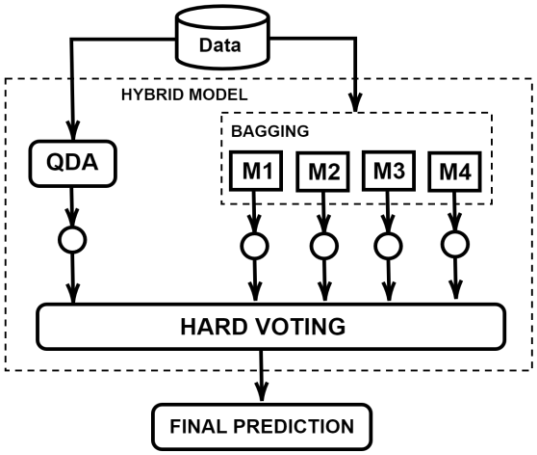


Figure 6. Architecture of the Hybrid Model

3.1.3. Linear discriminant analysis

Linear Discriminant analysis is also a dimensionality reduction method used to model the differences between the groups/classes. The higher dimension space is projected into the lower dimension space.

3.1.4. Quadratic discriminant analysis

QDA is quite related to linear discriminant analysis (LDA). QDA is a generative model and it assumes that every class follows a Gaussian distribution.

3.2. Ensemble techniques

3.2.1. Bagging

In parallel methods we fit the different considered learners independently from each other and, so, it is possible to train them concurrently. This approach is "bagging" ("bootstrap aggregation") that aims at producing an ensemble model that is more robust than the individual models composing it. The bagging approach is depicted in the Figure 7.

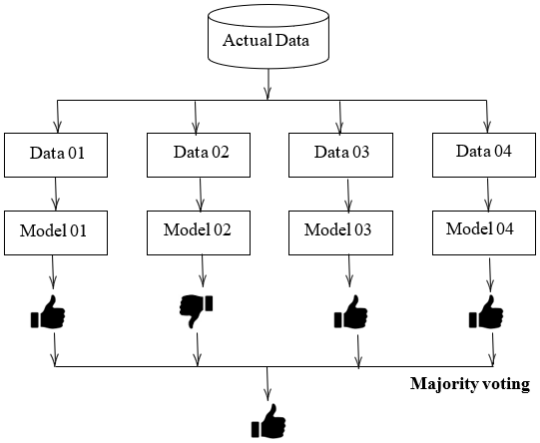


Figure 7. Bagging

3.2.2. Boosting

The idea is to fit models iteratively such that the training of models at a given step depends on the models fitted at the previous steps. This approach produces an ensemble model that is in general less biased than the weak learners that compose it. The boosting approach is depicted in the Figure 8.

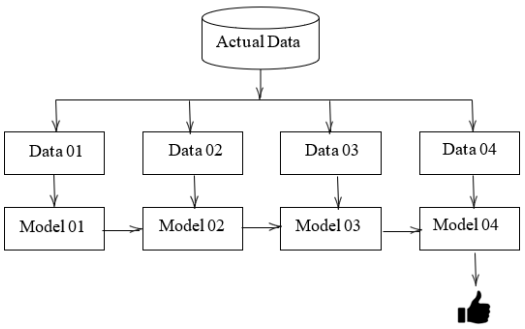


Figure 8. Boosting

3.2.3. Stacking

The idea is to learn several different weak learners and combine them by training a meta-model to output predictions based on the multiple predictions returned by these weak models. So, two things are defined in order to build our stacking model: The L learners to fit the model and the meta-model that combines them. Stacking is applied to single learners (first-level learners) constructed by different models. The meta-learner (second-level learner) is constructed by combining the training of different models to predict the output [Zhou 2012; Lateko *et al.* 2021]. The stacking approach is depicted in the Figure 9.

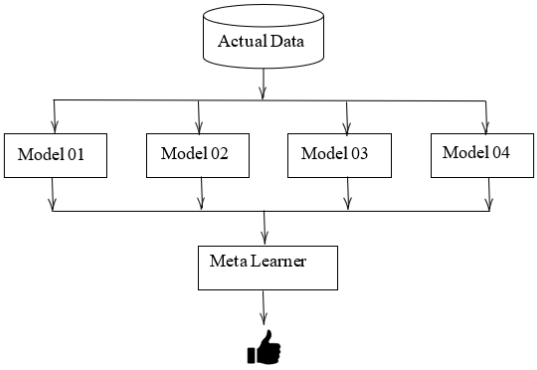


Figure 9. Stacking

3.3. Model Evaluation Metrics

3.3.1. Confusion Matrix

The Confusion Matrix is a matrix used for making out the performance of a classification model. The confusion matrix compares the actual values with the predicted values by the model.

3.4. Accuracy

Accuracy is the ratio of number of correct predictions to the total number of predictions. The closer the Accuracy of a model is to 1, the better the model.

In mathematical form, it may also be represented as follows:

3.5. Precision

$$\text{Accuracy}=\frac{(TN+TP)}{(TN+TP+FN+FP)} \quad (1)$$

Precision attempts to find the proportion of positive identifications that was actually correct. Eg: It refers to the rate of number of samples correctly predicted as drinkable out of all the samples classified as drinkable by the model.

3.6. Recall

$$\text{Precision}=\frac{TP}{(TP+FP)} \quad (2)$$

Recall refers to the proportion of actual positives that was identified correctly. Eg: It refers to the number of samples correctly predicted as drinkable out of all the samples that are actually drinkable.

3.7. F1-Score

$$\text{Recall}=\frac{TP}{(TP+FN)} \quad (3)$$

F1-score is the harmonic mean of precision and recall. F1-score ranges between 0 and 1. The closer it is to 1, the better the model.

$$F1\text{–}score=2*\frac{\text{Precision*Recall}}{(\text{Precision+Recall})} \quad (4)$$

3.8. ROC curve

ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. AUC stands for "Area under the ROC Curve."

4. Results

4.1. Statistical models

4.1.1. Binary korattur lake dataset

Table 1. Binary Class Korattur Lake Dataset Classification using Statistical Models

Statistical Algorithm	Accuracy
Principal Component Analysis	0.87
Hierarchical Clustering Analysis	0.53
Linear Discriminant Analysis	0.91
Quadratic Discriminant Analysis	0.95

Here, Quadratic Discriminant Analysis (QDA) is the most efficient algorithm with an accuracy of 95% as mentioned in the Table 1.

4.1.2. Binary Kaggle Dataset

Table 2. Binary Class Kaggle Dataset Classification using Statistical Models

Statistical Algorithm	Accuracy
Principal Component Analysis	0.88
Hierarchical Clustering Analysis	0.57
Linear Discriminant Analysis	0.88
Quadratic Discriminant Analysis	0.87

Here, Linear Discriminant Analysis (LDA) is the best algorithm with an accuracy of 88% as mentioned in the Table 2.

4.1.3. Three Class Korattur Lake Dataset

Table 3. Three Class Korattur Lake Dataset Classification using Statistical Models

Statistical Algorithm	Accuracy
Principal Component Analysis	0.75
Hierarchical Clustering Analysis	0.38
Linear Discriminant Analysis	0.92
Quadratic Discriminant Analysis	0.94

Here, Quadratic Discriminant Analysis (QDA) is the most efficient algorithm with an accuracy of 94% as mentioned in the Table 3.

4.1.4. Five Class Korattur Lake Dataset

Table 4. Five Class Korattur Lake Dataset Classification using Statistical Models

Statistical Algorithm	Accuracy
Principal Component Analysis	0.75
Hierarchical Clustering Analysis	0.17
Linear Discriminant Analysis	0.94
Quadratic Discriminant Analysis	0.97

Here, Quadratic Discriminant Analysis (QDA) is the most efficient algorithm with an accuracy of 97% as mentioned in the Table 4.

Overall, we can see that QDA is the most efficient algorithm from all the Statistical models.

So, it is chosen as the Statistical Algorithm to be used in the Hybrid Model.

4.2. Ensemble Models

4.2.1. Binary korattur lake dataset

Table 5. Performance metrics for binary Korattur Lake dataset using Ensemble Models

Ensemble Algorithm	Precision	Recall	F1-Score	Accuracy	Time (in seconds)
Bagging	1.0	1.0	1.0	1.0	0.464
Boosting	1.0	1.0	1.0	1.0	0.081
Stacking	1.0	1.0	1.0	1.0	2.916

Here, all three algorithms work best with an accuracy of 100% as mentioned in the Table 5.

4.2.2. Binary Kaggle Dataset

Table 6. Performance metrics for binary Kaggle dataset using Ensemble Models

Ensemble Algorithm	Precision	Recall	F1-Score	Accuracy	Time (in seconds)
Bagging	0.898	0.785	0.837	0.967	6.321
Boosting	0.7	0.047	0.089	0.88	0.114
Stacking	0.916	0.8	0.854	0.96	11.324

Here, the Bagging Algorithm works best with an accuracy of almost 97% as mentioned in the Table 6.

4.2.3. Three Class Korattur Lake Dataset

Table 7. Performance metrics for three class Korattur Lake dataset using Ensemble Models

Ensemble Algorithm	Precision	Recall	F1-Score	Accuracy	Time (in seconds)
Bagging	1.0	1.0	1.0	1.0	2.482
Boosting	0.99	0.99	0.99	0.99	0.101
Stacking	0.99	0.99	0.99	0.99	7.746

Here, the Bagging Algorithm works best with an accuracy of 100% as mentioned in the Table 7.

4.2.4. Five Class Korattur Lake Dataset

Table 8. Performance metrics for five class Korattur Lake dataset using Ensemble Models

Ensemble Algorithm	Precision	Recall	F1-Score	Accuracy	Time (in seconds)
Bagging	1.0	1.0	1.0	1.0	2.468
Boosting	1.0	1.0	1.0	1.0	0.165
Stacking	0.99	0.99	0.99	0.99	10.444

The performance of bagging and boosting is better compared to sacking approach which is given in the Table 8.

Here, both the Bagging and Boosting algorithms work best with an accuracy of 100%. Overall, Bagging model performs best. Hence, Bagging is chosen to be used in the Hybrid Model.

The time taken to predict the quality of water by each ensemble model is also tabulated. It is identified that boosting method takes least time compared to bagging and stacking. Stacking method takes lot of time as it has to construct the meta learner from the different base learners.

4.3. Hybrid Model

The Hybrid Model is the combination of both the best Statistical method - QDA and Ensemble method - Bagging. The combination of both QDA and Bagging was done using the voting classifier. The voting classifier is an ensemble classifier algorithm which trains various base models / estimators. The prediction is then done based on the combination of the findings of each base estimator.

4.3.1. Korattur Lake Dataset

The Hybrid Model performs with an accuracy of 100% on the Binary Class Korattur Lake Dataset.

4.3.2. Kaggle Dataset

The Hybrid Model performs with an accuracy of 96% on the Binary Class Kaggle Dataset.

4.3.3. Three Class Korattur Lake Dataset

The Hybrid Model performs with an accuracy of 100% on the three Class Korattur Lake Dataset.

4.3.4. Five Class Korattur Lake Dataset

The Hybrid Model performs with an accuracy of 100% on the five Class Korattur Lake Dataset.

4.4. Comparison

Bagging and Hybrid models performed the best for both binary and multi-class classification of Korattur Lake water data. While these models resulted with 96% and 97% for Bagging and Hybrid respectively for water dataset collected from Kaggle. QDA method resulted with 87% and ~94% to ~97% for water dataset from Kaggle and Korattur Lake (Figure 10).

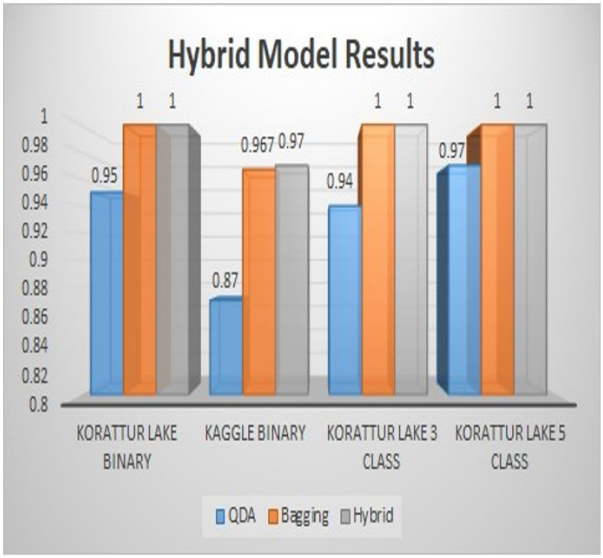


Figure 10. Accuracy comparison of the Ensemble, Statistical and Hybrid models across all the dastasets.

5. Conclusion

The evaluation and performance comparison of Statistical and Ensemble models were done and a Hybrid model was implemented, combining both. The results of the Hybrid model was compared with both the best performing Statistical and Ensemble models. According to the performance comparison of the three models, the Hybrid model is observed to be performing the best, irrespective of the dataset used. As a part of our future work, this research intends to expand its scope by introducing timestamp into the Binary, Three-Class and Five-Class Korattur lake water data, because the records of the dataset are ordered by time. This research intends to expand the scope of the project by implementing a time-series models such as : MA(Moving Average), ARIMA (Auto Regressive Integrated Moving Average), SARIMA (Seasonal Auto Regressive Integrated Moving Average) using the above timestamped datasets.

References

Abobakr Saeed Abobakr Yahya, Ali Najah Ahmed, Faridah Binti Othman, Rusul Khaleel Ibrahim, Haitham Abdulmohsin Afan, Amr El-Shafie, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, and Ahmed Elshafie. (2019). Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios. *Water*, **11(6)**.

Ali Omran Al-Sulttani, Mustafa Al-Mukhtar, Ali B. Roomi, Aitazaz Ahsan Farooque, Khaled Mohamed Khedher, and Zaher Mundher Yaseen. (2021). Proposition of new ensemble data-intelligence models for surface water quality prediction. IEEE Access, **9**, 108527–108541.

Al-Mahfoodh Najah, Ahmed El-Shafie, Othman A Karim, Othman Jaafar, and Amr El-Shafie. (2011). An application of different artificial intelligences techniques for water quality prediction. International Journal of Physical Sciences, **6**, 10.

Arshia Fathima, J Alamelu Mangai, and Bharat B Gulyani. (2014). An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques, International journal of river basin management, **12(4)**, 357–366.

Chenguang Song and Leihua Yao. (2022). A hybrid model for water quality parameter prediction based on ceemdan-ialo-lstm ensemble learning, Environmental Earth Sciences, **81(9)**, 1–14.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, **40**, 139-157.

Dipankar Ruidas, Subodh Chandra Pal, Towfiqul Islam, Abu Reza Md, and Asish Saha. (2022). Hydrogeochemical evaluation of groundwater aquifers and associated health hazard risk mapping using ensemble data driven model in a water scares plateau region of eastern india. Exposure and Health, pages 1–19

FARID HASSANBAKI GARABAGHI, Semra Benzer, and Recep Benzer. (2021). Performance evaluation of machine learning models with ensemble learning approach in classification of water quality indices based on different subset of features.

Gozen Elkiran, Vahid Nourani, and S.I. Abba. Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. Journal of Hydrology, 577:123962, 2019.

Jungsu Park, Woo Hyoung Lee, Keug Tae Kim, Cheol Young Park, Sanghun Lee, and Tae-Young Heo. (2022). Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Science of The Total Environment, **832**, 155070.

Jungsu Park. (2021). “the effect of input variables clustering on the characteristics of ensemble machine learning model for water quality prediction.”. Journal of Korean Society on Water Environment, **37(5)**, 335–343.

Lateko A. A., Yang H. T., Huang C. M., Aprillia H., Hsu C. Y., Zhong J. L. and Phương N. H. (2021). Stacking ensemble method with the RNN meta-learner for short-term PV power forecasting. Energies, **14(16)**, 4733.

Leizhi Wang, Zhenduo Zhu, Lauren Sassoubre, Guan Yu, Chen Liao, Qingfang Hu, and Yintang Wang. (2021). Improving the robustness of beach water quality modeling using an ensemble machine learning approach. Science of The Total Environment, **765**, 142760.

Li Z., Zhang C., Liu H., Zhang C., Zhao M., Gong Q. and Fu G. (2022). Developing stacking ensemble models for multivariate contamination detection in water distribution systems, Science of the Total Environment, **828**, 154284.

Li-Ming (Lee) He and Zhen-Li He. (2008). Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern california, usa. Water Research, **42(10)**, 2563–2573.

Lingbo Li, Jundong Qiao, Guan Yu, Leizhi Wang, Hong-Yi Li, Chen Liao, and Zhenduo Zhu. (2022). Interpretable tree-based ensemble model for predicting beach water quality. Water Research, **211**, 118078.

Mosavi A., Sajedi Hosseini F., Choubin B., Goodarzi M., Dineva A. A. and Rafiei Sardooi E. (2021). Ensemble boosting and bagging based machine learning models for groundwater potential prediction. Water Resources Management, **35**, 23–37.

Navideh Noori, Latif Kalin, and Sabahattin Isik. (2020). Water quality prediction using swat-ann coupled approach. Journal of Hydrology, **590**, 125220.

Ozgur Kisi, Meysam Alizamir, and AliReza Docheshmeh Gorgij. (2020). Dissolved oxygen prediction using a new ensemble method. Environmental Science and Pollution Research, **27(9)**, 9589–9603.

Pham Q.B. Saini G. *et al.* Abba S.I. (2020). Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index, Environmental Science and Pollution Research, **27**, 41524–41539.

Rahim Barzegar, Asghar Asghari Moghaddam, Jan Adamowski, and Bogdan Ozga-Zielinski. (2018). Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model, Stochastic environmental research and risk assessment, **32(3)**, 799–813.

Reddy, P., Talwar, P. B. and Gomathi, R. D. (2023). A Novel Model Using Multiple Bagging Ensemble Method for Measuring, Inferring and Predicting the Quality of Continuous Assessment Question Papers. In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Rodelyn Avila, Beverley Horn, Elaine Moriarty, Roger Hodson, and Elena Moltchanova. (2018). Evaluating statistical model performance in water quality prediction. Journal of Environmental Management, **206**, 910–919.

Rosaida Rosly, Mokhairi Makhtar, Mohd Khalid Awang, and Nordin Abdul. Comparison of ensemble classifiers for water quality dataset.

Sanghyun Park, Kyunghyun Kim, Changmin Shin, Joong-Hyuk Min, Eun Hye Na, and Lan Joo Park. (2020). Variable update strategy to improve water quality forecast accuracy in multivariate data assimilation using the ensemble kalman filter, Water Research, **176**, 115711.

Sani Isah Abba, Nguyen Thi Thuy Linh, Jazuli Abdullahi, Shaban Ismael Albrka Ali, Quoc Bao Pham, Rabiu Aliyu Abdulkadir, Romulus Costache, Van Thai Nam, and Duong Tran Anh. (2020). Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration. IEEE Access, 8:157218–157237.

Shuangyin Liu, Haijiang Tai, Qisheng Ding, Daoliang Li, Longqin Xu, and Yaoguang Wei. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. Mathematical and Computer Modelling, **58(3)**, 458–465, Computer and Computing Technologies in Agriculture 2011 and Computer and Computing Technologies in Agriculture.

Venkata Vara Prasad D, Lokeswari Y Venkataramana, P. Senthil Kumar, Prasannamedha G, Soumya K., and Poornema A.J. (2020). Water quality analysis in a lake using deep learning methodology: prediction and validation, International Journal of Environmental Analytical Chemistry, **0(0)**, 1–16.

Venkata Vara Prasad D, Lokeswari Y Venkataramana, P. Senthil Kumar, Prasannamedha G, Soumya K., and Poornema A.J. (2021). Prediction on water quality of a lake in chennai, india using deep learning algorithms. Desalination and Water Treatment, **218**, 44–51.

Venkata Vara Prasad D, P.Senthil Kumar, Lokeswari Y. Venkataramana, G. Prasannamedha, S. Harshana, S.Jahnavi Srividya, K. Harrinei, and Sravya Indraganti. (2021). Automating water quality analysis using ml and auto ml techniques. Environmental Research, **202**, 111720.

Xingguo Chen, Houtao Liu, Xiuying Xu, Luoyuan Zhang, Tianchi Lin, Min Zuo, Yichao Huang, Ruqin Shen, Da Chen, and Yongfeng Deng. (2021). Identification of suitable technologies for drinking water quality prediction: A comparative study of traditional, ensemble, cost-sensitive, outlier detection learning models and sampling algorithms. ACS ES&T Water, **1(8)**, 1676–1685.

Y Khan and SS Chai. (2017). Ensemble of ann and anfis for water quality prediction and analysis-a data driven approach. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), **9(2–9)**, 117–122.

Yunrong Xiang and Liangzhong Jiang. Water quality prediction using ls-svm and particle swarm optimization. pages 900–904.

Zengrui Huang, Wei Mao, Ming Chen, Qiang Wu, Boyue Xiong, and Wei Xu. (2019). An intelligent operation and maintenance system for power consumption based on deep learning. IOP Conference Series: Materials Science and Engineering, **486**, 012107.

Zhou Z.H. Combination Methods. (2012). In Ensemble Methods: Foundations and Algorithms; Taylor & Francis: Boca Raton, FL, USA; pp. 67–97.

Zilin Li, Chi Zhang, Haixing Liu, Chao Zhang, Mengke Zhao, Qiang Gong, and Guangtao Fu. (2022). Developing stacking ensemble models for multivariate contamination detection in water distribution systems. Science of The Total Environment, **828**, 154284.