Ensemble of ANN and ANFIS for Water Quality Prediction and Analysis - A Data Driven Approach

Y.Khan, S.S.Chai
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak, Kota Samarahan 94300, Malaysia.
yafra.khan@gmail.com

Abstract—**The consequences of un-clean water are some of the direst issues faced by humanity today. These concerns can be addressed efficiently if data is pre-analyzed and water quality is predicted before its effects occur. The aim of this research is to develop a novel ensemble of Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) models using averaging ensemble technique, producing improved prediction accuracy. Measurements of different water quality parameters have been used for predicting the overall water quality, applying ANN, ANFIS and ANN-ANFIS ensemble and their results have been compared. The data used in this study is obtained by USGS online repository for the year of 2015, with a 30-minutes time interval between measurements. Root Mean Squared Error (RMSE) has been used as the main performance measure. The results depict a significant improvement in the Ensemble ANN-ANFIS model (RMSE: 0.457) as compared to both the ANN model (RMSE: 2.709) and the ANFIS model (1.734). The study concludes that the ensemble of ANN and ANFIS model shows significant improvement in prediction performance as compared to the individual models. The research can prove to be beneficial for decision making in terms of water quality improvement.**

Index Terms—**Water Quality Prediction; Artificial Neural Networks; Adaptive Neuro-Fuzzy Inference System; Ensemble Learning; Machine Learning.**

I. Introduction

The contamination of natural water resources is quite rampant due to its wide availability. This contamination is the result of various factors including poor sanitation infrastructure and lack of awareness [1]. This engenders a dire need for adopting innovative approaches and techniques for water quality prediction before its consequences arise and take the precautionary actions. Water quality can be evaluated by either a single parameter for a specific use or by multiple Water Quality (WQ) parameters. In case of multiple WQ parameters, a Water Quality Index (WQI) is used, which is a numerical representation of the quality of a water resource covering various significant water quality parameters in connection with a set of water quality standards [2].

For carrying out the analysis and prediction of water quality, various studies have proposed and implemented different methodologies [3]. One such study [4] proposes Reasoning Based Expert System (RBES) to compare the water quality parameters with the industry standards from the knowledge base to make a decision. Besides that, time-series analysis techniques like Auto-Regressive Integrated Moving Average (ARIMA) have been widely used in this regard [4][6]. More recently, Support Vector Machine (SVM) has been applied in water quality prediction scenario [7] to predict the concentration of one parameter in water based upon the values of other water quality parameters.

Despite improving results in the above mentioned techniques, following few points need to be considered when selecting a suitable technique for water quality prediction: a) Mapping input-output data in case of water quality dataset becomes very complex due to non-linear nature of water quality dataset with linear modeling approaches [8] [9] (b) Prediction accuracy and model simplicity needs to be considered (c) Simplified interpretation of input-output relationship in order to deal with uncertainties (d) Combining multiple models improves generalization and diversity of the model.

Artificial Neural Network (ANN) is one technique that has proved to be effective in not only describing nonlinear input-output relationship of complex datasets, but also in providing strong model flexibility [10]. ANN has been applied successfully for other complex prediction scenarios like groundwater level prediction [11] and wind speed forecasting [12]. In case of water quality prediction, Gazzaz et.al. [2] use Multi-Layer Perceptron (MLP) Neural Network to predict WQI based upon certain WQ parameters. The result in terms of RMSE turns out to be effective. On the downside, ANN is a black-box approach, hence the model simplicity is compromised and uncertainty is not effectively dealt with. Adaptive Neuro-Fuzzy Inference System (ANFIS) has been found to be an effective approach in this regard, using the interpretability aspect of fuzzy inference while retaining the benefits of ANN [11]. ANFIS has been found to be suitable in modeling complex datasets like that of hydrological applications. One such study applied ANFIS for prediction of oily wastewater microfiltration permeate volume and was found to be a reliable approach [13]. Similarly, Talebizadeh and Moridnejad [14] carried out a comparison of ANN and ANFIS in forecasting lake level fluctuations, where ANFIS turned out to be superior than ANN in terms of efficiency. In case of water quality prediction, ANFIS has been applied for Biochemical Oxygen Demand (BOD) prediction based upon other WQ parameters as inputs [15]. This study shows a significant accuracy for different input combinations, with MSE between 1.2 and 2.5.

Despite prediction model improvements, increase in model accuracy while avoiding over-fitting is still a challenge for most researchers. According to recent researches, model performance can be significantly improved if an appropriate hybrid of multiple models is used for forecasting and prediction than using a single model in this regard [16][6]. Ensemble learning refers to a process of combining multiple predictors in order to boost the model performance. It uses a combination or a committee of relatively “weak” learners to achieve a better performance [17]. There are different methods of creating ensembles, depending upon the requirement. Ensemble Learning has been used in various applications like forecasting energy consumption [18] and classification of cancer [19]. However, the use of Ensemble Learning for water quality prediction is a fairly recent research area. A significant study in this regard proposes a homogeneous ensemble of ANN models for water quality parameter prediction [9], selecting an optimal model for each WQ parameter.

The central theme of this study is to propose a comprehensive procedure of analyzing and predicting the quality of water in case of a particular region using certain water quality parameters and seek to improve prediction accuracy. An ensemble technique based upon the model averaging technique has been proposed to combine the ANN and ANFIS models and the results are computed and analyzed. The comparison between the individual ANN and ANFIS models with the ensemble model has been carried out and outcome is analyzed in terms of performance [3].

II. Materials and Methods

Data collection and analysis plays an integral role in the effectiveness of prediction models. The data variation and richness makes sure the prediction models take into account the several aspects of the process [20]. Due to the lack of detail and inconsistent observations of most water monitoring organization, this research opts for one of the most comprehensive and reliable water quality data resource available today in order to acquire the data. The U.S. Geological Survey's (USGS) data repository called National Water Information System (NWIS) has been selected for the acquisition of sample data for this study. NWIS is an open data resource allowing the acquisition, processing and storage of water quality data across the U.S. The selected study area comprises of a water channel in the Island Park village, located around the New York County of South-Western Nassau with the Latitude 40°36'31.8", Longitude 73°39'22.0" as shown in Figure 1.

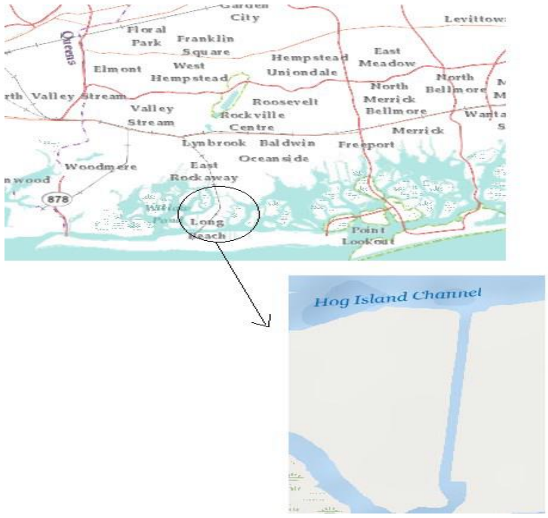

Figure 1: Area covered in the Hog Island Channel Monitoring Station

The methodology and model architecture components have been explained below:

A. *Water Quality Index (WQI)*

This study uses WQI as a means of estimating the overall quality of water resource and is treated as the model output. The WQI has been acquired using the formula developed by National Sanitation Foundation (NSF) [21]. This WQI is calculated by the following formula:

$$WQI = \sum_{i=1}^{n} S_i \,. W_i \tag{1}$$

where S_i and W_i are the sub-index and weight for the ith WQ parameter and n is the number of parameters. Initially there are nine parameters selected for this study, however, the parameters are reduced to five based upon the sensitivity analysis technique in [22], making the process faster, less complex and more cost effective [23]. The initial water quality parameters with weights are detailed in Table 1. After sensitivity analysis, the remaining parameters are Specific Conductance (SC), Dissolved Oxygen (DO), Nitrate (Ni), pH and Chlorophyll (Chl).

Table 1
WQ Parameters and their Weights

Factor	Unit	Weight
DO	% Saturation	0.17
pH	pH unit	0.16
Chlorophyll	μg/L	0.11
Specific Conductance	μS/cm	0.11
Temperature	°C	0.10
Salinity	PSE	0.10
Nitrate	mg/L	0.10
Turbidity	FNU	0.08
Water Surface Elevation	Ft.	0.07

B. *Artificial Neural Networks (ANN)*

In order to model and classify complex datasets, ANN has proved to be an effective methodology, particularly for datasets related to environmental processes. The core strength if ANN as a prediction model is that it caters to the non-linear relationship of the input and output water quality datasets [14]. Its basic structure comprises of an input layer, hidden layer and output layer each consisting of nodes. A general feed-forward and back-propagation Neural Network consists of three layers, i.e. one input, one hidden and one output layer. There are two processes involved, namely feed forward and back-propagation. In the feed forward process, the initial weights are multiplied by the inputs and the subsequent value moves to the next layer, till it arrives at the output layer, shown by following equation:

$$z_i = \sum_{j=1}^{m} w_{ij} x_{ij} \tag{2}$$

where $\boldsymbol{w_{ij}}$ represents the weight moved from ***j***th input to the ***i***th node, $\boldsymbol{x_{ij}}$ depicts the input while $\boldsymbol{z_i}$ denotes the resultant summation of outputs of the ***i***th node. After this step, the error value is calculated through the back-propagation process, by determining the difference between predicted value and target value. It starts backwards from the output layer to the input layer[6]. The difference is represented by the symbol $\boldsymbol{\delta(l)_j}$, showing the error of node *j* in layer *l*. The error term for a training set (x_j,y_j) is shown by:

$$\boldsymbol{\delta(l)_j = z_j - y_j} \tag{3}$$

The previous process goes on repeatedly, while adjusting the weights, until convergence.

For this study, input and target data was divided into training data (70%) and testing (30%). The input consists of water quality parameter values and targets are the calculated WQI, whereas the output is the predicted WQI. The input layer, with reduced inputs, consists of five WQ variables; so the model has a 5:10:1 architecture, implying that there are five inputs (WQ parameters), ten hidden nodes and one output (WQI) (Figure 2). A feed-forward Neural Network has been used with the training algorithm of Scaled Conjugate Gradient (SCG) and the activation function for the hidden layer is Log Sigmoid (LSIG), while that for the output layer is Linear.

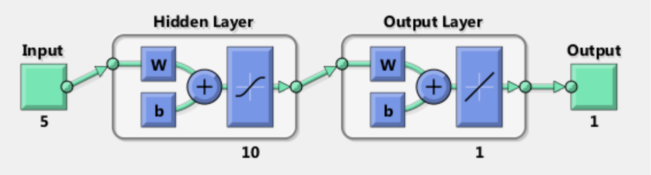

Figure 2: Schematic Diagram of ANN Architectur

C. Adaptive Neuro-Fuzzy Inference System (ANFIS)

Adaptive Neuro-Fuzzy Inference System (ANFIS) is a Multi-Layer Feed-Forward network using learning algorithm of neural network and fuzzy logic in order to map inputs with outputs [24]. ANFIS uses Takagi–Sugeno type Fuzzy Inference System (FIS), where the output of each fuzzy rule can be a linear combination of input variables plus a constant term. ANFIS generally uses two types of learning algorithms i.e. the back propagation and hybrid learning. The back-propagation learning is used similar to that of back propagation in ANN. The hybrid learning consists of a combination of back propagation and least squares method. It uses Least Squares Method (LSM) for forward-passing and Gradient Descent for back propagation in the training process of ANFIS [25].

This study uses hybrid learning algorithm for ANFIS as it is much faster to converge than the conventional back propagation method [26]. For ANFIS implementation, first the input and target data is loaded and randomly divided into training (70%) and testing data (30%). The architecture consists of five inputs (WQ parameters) and one output (WQI). The Fuzzy Inference System (FIS) is then generated using Fuzzy C-Means clustering with number of clusters set to 15. After that, the input and output membership functions (mf) are created, using Gaussian membership function (gaussmf). The ANFIS architecture consists of five inputs, fifteen membership functions, fifteen fuzzy rules and one output.

D. Ensemble Learning

The models based upon multiple learners have been shown to perform better than models with single learners, especially when dealing with complex datasets [27]. The branch of machine learning dealing with multiple homogenous or heterogeneous models is collectively termed as ensemble learning. The intuition is to use a combination or a committee of relatively “weak” learners to achieve a better performance [17]. The basic component of ensemble learning is a base learner which is created with a base learning algorithm. The generalized ensemble equation is given by:

$$H(x) = \sum_{i=1}^{n} h_i(x) \tag{4}$$

where *h(x)* denotes a single predictor and *H(x)* denotes an ensemble with *n* total number of predictors. There are different methods of creating ensembles, including bagging, boosting, majority voting and averaging. This study implements the model averaging approach for combining the ANN and ANFIS predictors.

The averaging ensemble technique first trains the ANN and ANFIS model and tests them separately. It then generates the average value of the ANN and ANFIS training output for each example. Finally, the average output is tested against the test set. The weighted model average is given by:

$$\Phi(x) = \frac{1}{M} \sum_{m=1}^{M} P_m \tag{5}$$

where *M* is the number of learners and *P* denotes the function *f(x)*. The implementation is in the following steps:

1. For each predictor, P:
    - Input and Target data is loaded, with 70% training data and 30% test data.
    - Algorithm is applied through the training data and error is generated. Prediction output is stored.
    - After training, algorithm is tested against the tested data and error is calculated. Its prediction output is stored.
2. Each generated training prediction from predictor P1 and P2 is averaged and the final training output is stored.
3. Each generated testing prediction from predictor P1 and P2 is averaged and the final testing output is stored.
4. Training and testing errors of the final outputs are calculated.

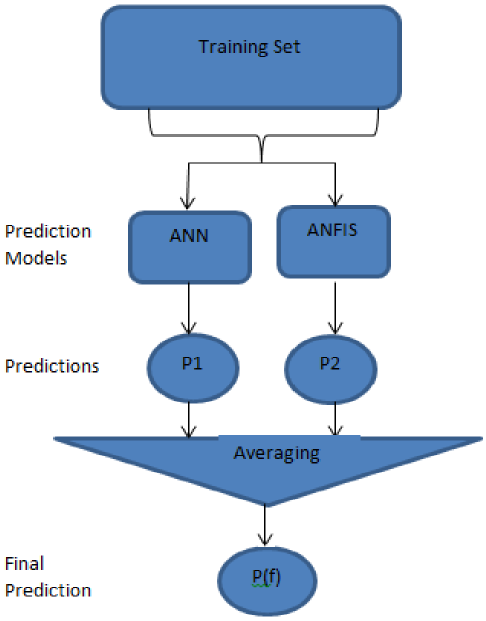

Figure 3: Ensemble Implementation

III. RESULTS AND DISCUSSION

Three tests have been performed to analyze the prediction of overall water quality based upon the water quality factors. Firstly, ANN feed-forward model with the training algorithm of Scaled Conjugate Gradient (SCG) is applied. As depicted in Figures 4(a), 4(b) and Table 2 (WQI-ANN-Opt Model), the significant aspects of the output are epochs (iteration), Regression, MSE and RMSE of both training and testing data. The number of epochs, which is the number of iterations it takes for the model to converge, is 96, with best validation performance at epoch 90. In this case, the training RMSE is 2.709. On the other hand, the Regression plot of training, validation and testing data (Figure 4(b)) shows the function fit through the scatter plot with observed WQI on x-axis and predicted WQI on y-axis. The closer the regression value is to 1, the better the function fits. As seen in the graph, most of the data points are close to the regression line, with few outliers. The regression values of training and testing data are 0.972 and 0.973 respectively.

The second test uses ANFIS model to predict water quality. WQI-ANFIS model consists of 5 inputs and one output, depicting the predicted WQI. The plot of Observed WQI and Predicted WQI of both training and testing data in Figure 5(a) shows clearly that most of the data points fit to the function, with very few outliers. The R values of 0.988 and 0.986 show a balanced function fit. Furthermore, Figure 5(b) and Figure 5(c) show the error plots of training and testing data respectively. The MSE and RMSE of the training data are 3.007 and 1.734 respectively. The MSE and RMSE values imply that the prediction accuracy with the ANFIS-WQI model has improved considerably as compared to the ANN-WQI model. The same can be observed in terms of test data.

The third and final test implements the ANN-ANFIS ensemble of model averaging to further improved prediction performance. The error distribution plot shows that most of the errors are concentrated near zero. It can be seen from the error plot that ANN-ANFIS ensemble ensures a smooth error range, likely to be converged quickly with 70 epochs (Table 4). As seen from the training performance figure (Figure 6(a)), the training accuracy of MSE 0.161 is achieved while RMSE is 0.401. The MSE and RMSE of testing turns out to be 0.292 and 0.540 respectively (Figure 6(b)). When training and testing errors are analysed, it can be seen that the error values do not deviate much from each other, showing good generalization ability. The R value of 0.904 depicts a balanced function fit (Figure 6(c)).

It should be noted that in the datasets like those of hydrological systems, the RMSE value is not necessarily in agreement with Correlation Coefficient (R) as inversely proportional. In other words, a model is depicted as a good predictor by R value even when it is not, as noted by [28] and [29]. Hence RMSE is treated as a better indicator of model performance.

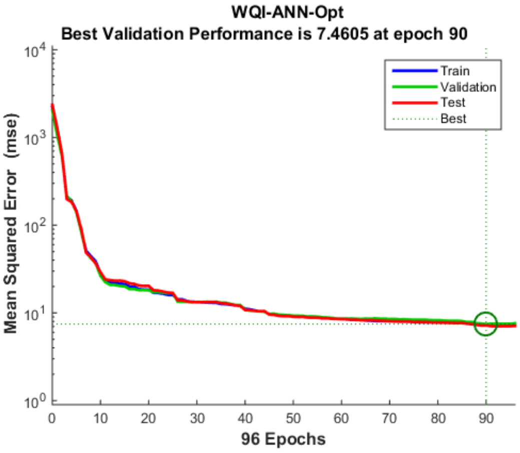

Figure 4(a): Mean Squared Error for ANN

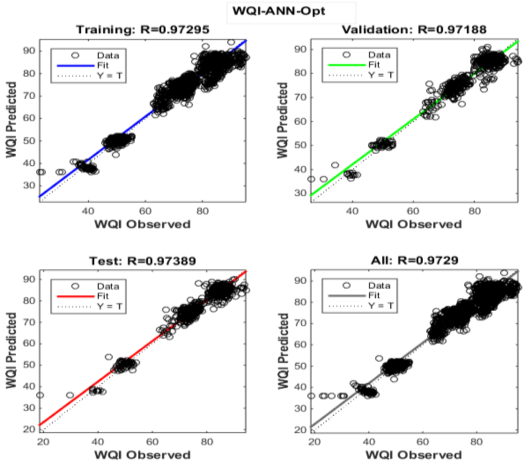

Figure 4(b): Regression for ANN

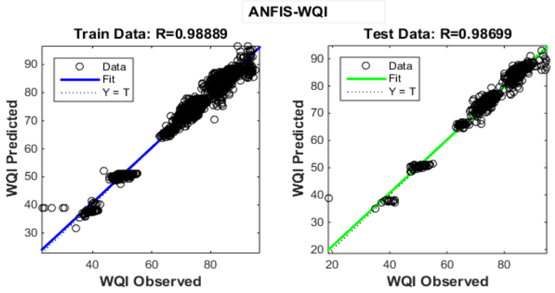

Figure 5(a): Regression for ANFIS-WQI Model

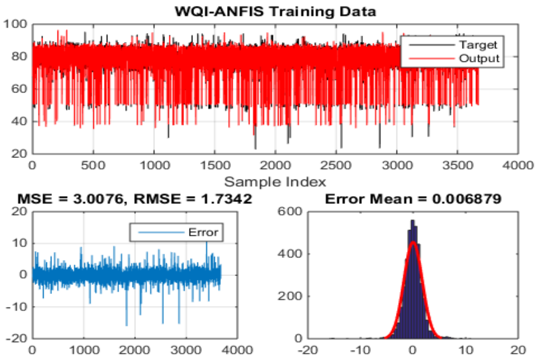

Figure 5(b): Training Error for ANFIS-WQI Model

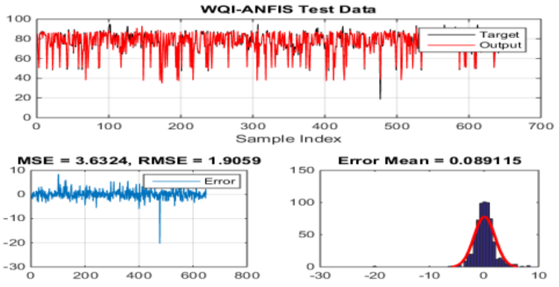

Figure 5(c): Testing Error for ANFIS-WQI Model

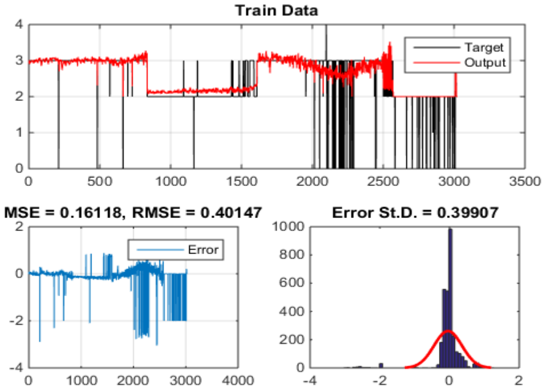

Figure 6(a): Training MSE for ANN-ANFIS Ensemble Model

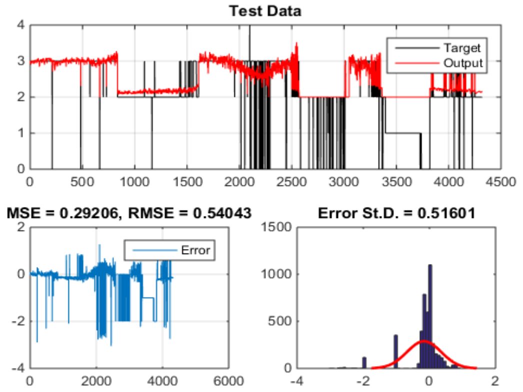

Figure 6(b): Training MSE for ANN-ANFIS Ensemble Model

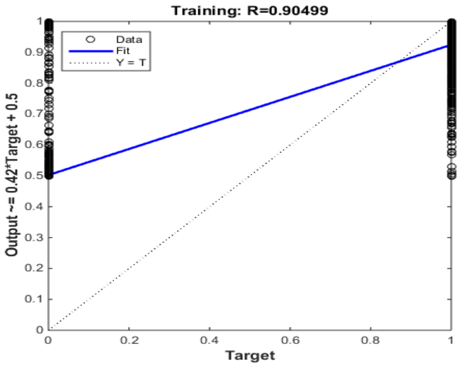

Figure 6(c): Training MSE for ANN-ANFIS Ensemble Model

Table 2
Performance Measures for ANN model

		Training Data			Testing Data		
Variables	Model	R	MSE	RMSE	R	MSE	RMSE
Reduced Variables	ANN-WQI-Opt	0.972	7.341	2.709	0.973	7.197	2.682

Table 3
Performance Measures for ANFIS Model

		Training Data			Testing Data		
Model	Epochs	R	MSE	RMSE	R	MSE	RMSE
ANFIS-WQI	100	0.988	3.007	1.734	0.986	3.632	1.905

Table 4
Performance Measures for ANN-ANFIS Ensemble Model

		Training Data			Testing Data		
Model	Epochs	R	MSE	RMSE	R	MSE	RMSE
ANN-ANFIS Ensemble	70	0.904	0.161	0.401	0.986	0.292	0.540

IV. Conclusion

This paper seeks to predict the quality of water in terms of Water Quality Index (WQI), with water quality parameters as inputs. The data for this water quality has been obtained from an online repository of USGS, called National Water Information System (NWIS). The data comprises of the measurements of water quality parameters with 30-minute time interval from the year of 2015. The study area is a channel situated in the State of New York. A hybrid ensemble of Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) has been used in order to determine the prediction accuracy and test the model for water quality application. The model performance is compared with the individual ANN and ANFIS models in terms of prediction accuracy and model diversity. To evaluate the model performance, Correlation Coefficient (R), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used.

The results of the tests performed give an idea about the model performance. A comparison between the ensemble model and individual proves the ANN-ANFIS Ensemble model to be more accurate, with the prediction accuracy indicating much improved values (training RMSE=0.401, testing RMSE=0.540), as compared to individual ANN model (training RMSE=2.708, testing RMSE=2.682) and ANFIS model (training RMSE=1.734, testing RMSE=1.905). As this study depicts a better result with a hybrid ensemble machine learning technique, more hybrid models need to be devised for further improvements in prediction performance. In addition to further improvements in model performance, a more user-centric approach should be adopted towards addressing the water quality issues, by involving all the significant stakeholders, using user-friendly tools and an interactive platform so that the solution truly benefits the target users.

REFERENCES

[1] P. Zeilhofer, "GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil," *Cad. Saúde ...*, vol. 23, no. 4, pp. 875–884, 2007.

[2] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors," *Mar. Pollut. Bull.*, vol. 64, no. 11, pp. 2409–2420, 2012.

[3] Y. Khan and C. Soo See, "Predicting and Analyzing Water Quality using Machine Learning : A Comprehensive Model," in *2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, 2016, pp. 1–6.

[4] Y. Wang, Y. Wang, M. Ran, Y. Liu, Z. Zhang, L. Guo, Y. Zhao, and P. Wang, "Identifying Potential Pollution Sources in River Basin via Water Quality Reasoning Based Expert System," *2013 Fourth Int. Conf. Digit. Manuf. Autom.*, pp. 671–674, 2013.

[5] A. Tizro, M. Ghashghaie, P. Georgiou, and K. Voudouris, "A r w w," vol. 1, pp. 43–52, 2014.

[6] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data," *Appl. Soft Comput.*, vol. 23, no. January 2016, pp. 27–38, 2014.

[7] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea.," *Sci. Total Environ.*, vol. 502, pp. 31–41, Jan. 2015.

[8] C. Min, "An Improved Recurrent Support Vector Regression Algorithm for Water Quality Prediction," vol. 12, pp. 4455–4462, 2011.

[9] S. E. Kim and I. W. Seo, "Artificial Neural Network ensemble modeling with conjunctive data clustering for water quality prediction in rivers," *J. Hydro-environment Res.*, Apr. 2015.

[10] S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction," *Environ. Earth Sci.*, vol. 71, no. 7, pp. 3147–3160, 2013.

[11] Y. Gong, Y. Zhang, S. Lan, and H. Wang, "A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida," *Water Resour. Manag.*, pp. 375–391, 2015.

[12] O. Baghirli, "Comparison of Lavenberg-Marquardt, Scaled Conjugate Gradient And Bayesian Regularization Backpropagation Algorithms for Multistep Ahead Wind Speed Forecasting Using Multilayer Perceptron Feedforward Neural Network," *Dissertation*, no. June, p. Uppsala University, 2015.

[13] A. Rahimzadeh, F. Z. Ashtiani, and A. Okhovat, "Application of adaptive neuro-fuzzy inference system as a reliable approach for prediction of oily wastewater microfiltration permeate volume," *J. Environ. Chem. Eng.*, vol. 4, no. 1, pp. 576–584, 2016.

[14] M. Talebizadeh and A. Moridnejad, "Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ANN and ANFIS models," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4126–4135, 2011.

[15] A. A. M. Ahmed and S. M. A. Shah, "Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River," *J. King Saud Univ. - Eng. Sci.*, p. , 2015.

[16] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5–6, pp. 781–789, 2005.

[17] H. Daume, "Ensemble Methods," in *A course in machine learning*, 2012, p. 189.

[18] S. Barak and S. S. Sadegh, "Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm," *Int. J. Electr. Power Energy Syst.*, vol. 82, pp. 92–104, 2016.

[19] S. Nagi and D. K. Bhattacharyya, "Classification of microarray cancer data using ensemble approach," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 2, no. 3, pp. 159–173, 2013.

[20] The Environmental and Protection Agency, "Parameters of water quality," *Environ. Prot.*, p. 133, 2001.

[21] M. Wills and K. N. Irvine, "Application of the National Sanitation Foundation Water Quality Index in the cazenovia Creek, MY, Pilot Watershed Management Project," *Middle States Geogr.*, pp. 95–104, 1996.

[22] H. Juahir, M. A. Zali, A. Retnam, S. M. Zain, M. F. Kasim, B. Abdullah, and S. B. Saadudin, "Sensitivity analysis for water quality index (WQI) prediction for kinta river, Malaysia," *World Appl. Sci. J.*, vol. 14, no. SPL ISS 1, pp. 60–65, 2011.

[23] N. Snchez-Marono and A. Alonso-Betanzos, *Feature selection based on sensitivity analysis*. 2007.

[24] H. Yan, Z. H. Zou, and H. W. Wang, "Adaptive neuro fuzzy inference system for classification of water quality status," *J. Environ. Sci.*, vol. 22, no. 12, pp. 1891–1896, 2010.

[25] C. Loganathan and K. V Girija, "Hybrid Learning For Adaptive Neuro Fuzzy Inference System," vol. 2, no. 11, pp. 6–13, 2013.

[26] C. Loganathan and K. V Girija, "Investigations on Hybrid Learning in ANFIS," *Int. J. Eng. Res. Appl.*, vol. 4, no. 10, pp. 31–37, 2014.

[27] P. Kazienko, E. Lughofer, and B. Trawiński, "Hybrid and ensemble methods in machine learning J.UCS special issue," *J. Univers. Comput. Sci.*, vol. 19, no. 4, pp. 457–461, 2013.

[28] D. R. Legates and G. J. McCabe Jr., "Evaluating the Use of 'Goodness of Fit' Measures in Hydrologic and Hydroclimatic Model Validation," *Water Resour. Res.*, vol. 35, no. 1, pp. 233–241, 1999.

[29] C. Willmott, "Some comments on the evaluation of model performance," *Bulletin of the American Meteorological Society*, vol. 63, no. 11. pp. 1309–1313, 1982.