EFFICIENT WATER QUALITY ANALYSIS & PREDICTION USING MACHINE LEARNING

LITERATURE SURVEY

This research explores the methodologies that have been employed to help solve problems related to water quality. Typically, conventional lab analysis and statistical analysis are used in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem. Local research employing lab analysis helped us gain a greater insight into the water quality problemin Pakistan. In one such research study, Daud et al. [5] gathered water samples from different areas of Pakistan and tested them against different parameters using a manual lab analysis and found a high presence of E. coli and fecal coliform due to industrial and sewerage waste. Alamgir et al. [6] tested 46 different samples from Orangi town, Karachi, using manual lab analysis and found them to be high in sulphates and total fecal coliform count. After getting familiar with the water quality research concerning Pakistan, we explored research employing machine learning methodologies in the realm of water quality. When it comes to estimating water quality using machine learning, Shafi et al. [7] estimated water quality using classical machine learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbors (kNN), with the highest accuracy of 93% with Deep NN. The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards (Available online at URL https://www.who.int/airpollution/guidelines/en/). Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality. Ahmad et al. [8] employed single feed forward neural networks and a combination of multiple neural networks toestimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a little immoderate in terms of an inexpensive real time system, given the price of the parameter sensors. Sakizadeh [9] predicted the WQI using 16 water quality parameters and

ANN with Bayesian regularization. His study yielded correlation coefficients between the observed and predicted values of 0.94 and 0.77, respectively. Abyaneh [10] predicted the chemical oxygen demand (COD) and the biochemical oxygen demand (BOD) using two conventional machine learning methodologies namely, ANN and multivariate linear regression. They used four parameters, namely pH, temperature, total suspended solids (TSS) and total suspended (TS) to predict the COD and BOD. Ali and Qamar [11] used the unsupervised technique of the average linkage (within groups) method of hierarchical clustering to classify samples into water quality classes. However, they ignored the major parameters associated with WQI during the learning process and they did not use any standardized water quality index to evaluate their predictions. Gazzaz et al. [4] used ANN to predict the WQI with a model explaining almost 99.5% of variation in the data. They used 23 parameters to predict the WQI, which turns out to be quite expensive if one is to use it for an IoT system, given the prices of the sensors. Rankovic et al. [12] predicted the dissolved oxygen (DO) using a feedforward neural network (FNN). They used 10 parameters to predict the DO, which again defeats the purpose if it has to be used for a real-time WQI estimation with an IoT system. Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough. The proposed methodology improves on these notions and the methodology being followed is depicted in Figure 1. Water 2019, 11, x FOR PEER REVIEW 3 of 14 three parameters and comparing them to standardized values is quite a limitation when predicting water quality. Ahmad et al. [8] employed single feed forward neural networks and a combination of multiple neural networks to estimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a little immoderate in terms of an inexpensive real time system, given the price of the parameter sensors. Sakizadeh [9] predicted the WQI using 16 water quality parameters and ANN with Bayesian regularization. His study yielded correlation coefficients between the observed and predicted values of 0.94 and 0.77, respectively. Abyaneh [10] predicted the chemical oxygen demand (COD) and the biochemical oxygen demand (BOD) using two conventional machine learning methodologies namely, ANN and multivariate linear regression. They used four

parameters, namely pH, temperature, total suspended solids (TSS) and total suspended (TS) to predict the COD and BOD. Ali and Qamar [11] used the unsupervised technique of the average linkage (within groups) method of hierarchical clustering to classify samples into water quality classes. However, they ignored the major parameters associated with WQI during the learning process and they did not use any standardized water quality index to evaluate their predictions. Gazzaz et al. [4] used ANN to predict the WQI with a model explaining almost 99.5% of variation in the data. They used 23 parameters to predict the WQI, which turns out to be quite expensive if one is to use it for an IoT system, given the prices of the sensors. Rankovic et al. [12] predicted the dissolved oxygen (DO) using a feedforward neural network (FNN). They used 10 parameters to predict the DO, which again defeats the purpose if it has to be used for a real-time WQI estimation with an IoT system. Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough. The proposed methodology. Ahmed et al. [4] have used the supervised machine learning algorithms in order to assess the water quality index (WQI), where an individual index was used to summarize the overall quality of water, and water quality class (WQC). Their suggested techniques and the gradient boosting with a learning rate of 0.1 and polynomial regression with a degree of 2 has predicted the WQI most effectively, and that WQI was subsequently determined with a mean absolute error (MAE) of 1.9642 and 2.7273. In this instance, the MLP, which has the configuration of (3, 7), has the highest classification accuracy of 85.07%. Wang et al. [6] have proposed a two-layered model stacking approach for predictive modeling of beach water quality. The five most frequently used methods (partial least square, sparse partial least square, random forest, Bayesian network, akhand linear regression) are integrated into a machine learning model that is then used to generate the final forecast. In this case, the model stacking technique was applied to three different beaches around eastern Lake Erie, New York, USA, and compared to all five basis models. After analysis, the model stacking strategy performed better than all of the base models. Year-over-year, stacking model accuracy scores were constantly at or near the top of the rankings, with a year-on-year accuracy average of 78%, 81%, and 82.3% at the three tested beaches. Sillberg et al. [7] have developed a machine learning-based approach integrating attribute-realization (AR) and support vector machine (SVM) algorithm to classify the Chao Phraya River's water quality. The AR has

determined the most significant factors to improve the river's quality using the linear function. In the categorization, the most contributing characteristics were: NH3 -N, TCB, FCB, BOD, DO, and Sal, boosting the contributed values in the range of 0.80–0.98, vs 0.25–0.64 for TDS, Turb, TN, SS, NO3-N, and Cond. The SVM linear method has enabled the best classification results represented as the accuracy of 0.94, a precision average of 0.84, recall average of 0.84, and F1-score average of 0.84. The validation showed that AR-SVM was a powerful method to identify river water quality with 0.86–0.95 accuracy when applied to three to six characteristics. Yilma et al. [8] have used an artificial neural network to simulate the Akaki River's WQI. The twelve water quality indicators from 27 dry and wet season sample locations were utilized to calculate the index. Except for one upstream location, all forecast results have shown low water quality. Here, the number of hidden layers (2–20), hidden layer neurons (5, 10, 15, 20, 25), transfer, training, and learning functions were used to train and verify the neural network model through
12 inputs and one output. Their study has revealed that an artificial neural network with eight hidden layers and 15 hidden neurons accurately predicted the WQI with an accuracy of 0.93. Bui et al. [9] have developed a random tree and bagging (BA-RT) hybrid machine learning method. Their research has tested four standalone (RF, M5P, RT, and REPT) and 12 hybrid data-mining algorithms (hybrids of the standalone with bagging, CVPS, and RFC) for forecasting monthly WQI in a humid climate in northern Iran. To forecast IRAQIs, they found that fecal coliform and total solids had the largest and least impact. Here, the optimal input combinations have differed across algorithms but the variables with poor correlations have performed worse. The Hybrid algorithms have improved their prediction power of several of the standalone models, but not all, and the Hybrid BA-RT has outperformed the other models by achieving R^2 0.941 using a 10-fold cross-validation technique, outdoing 15 standalone and hybrid algorithms. Ding et al. [10] have designed a hybrid intelligent method that combines Principal Component Analysis (PCA), Genetic Algorithm (GA), and Back Propagation Neural Network (BPNN) techniques for predicting river water quality. In this study, 23 different water quality indicator variables were utilized, each of which has a complicated non-linear connection to water quality. In this case, PCA has significantly increased the training speed of follow-up algorithms, while GA has optimized the parameters of BPNN. The average prediction rates for non-polluted and polluted water quality were 88.9% and 93.1% respectively, while the worldwide

prediction rate was around 91%, according to the results. Azad et al. [11] have utilized the three evolutionary algorithms including, GA, DE, and ACO_R in order to optimize the performance of adaptive neuro-fuzzy inference system (ANFIS) for water quality metrics prediction. These algorithms have been integrated with the ANFIS to predict the EC, SAR, and THE water quality metrics. Based on their research, the ANFIS-DE model, with an R^2 of 0.98 and an RMSE of 73.03, as well as a MAPE of 5.16, was the most accurate in predicting EC and TH in the test stage. Furthermore, the ANFIS-DE and ANFIS-GA models have shown the greatest performance in SAR (R^2 = 0.95, 0.91; RMSE = 0.43, 0.37; MAPE = 13.43, 13.72) prediction in a test stage. It has been shown that ANFIS is capable of producing the best results in the training stage with respect to water quality indicators. Zhang et al. [12] have improved a hybrid artificial neural network (HANN) model by the genetic algorithm (GA) for the prediction of drinking water treatment plants in china. The model has trained, validated, and has been continually validated using monthly data from 45 DWTPs across China that comprises eleven input variables for water quality and operational performance. The HANN model has shown better ability and consistency in forecasting the total water output of DWTPs in combination with the water quality and operational factors. Their prediction shows that the HANN model has improved its performance from 0.71 to 0.93 (R^2) by increasing the training data provided, as shown by the fact that the model has the ability to grow to the greatest level of performance.