Ideation phase

Modeling the quality of water resources is vitally important for water scheduling and

management. In the past, scientists regularly sampled the water in water quality monitoring

stations and assessed the components in the water sample in a lab. However, this process takes a

long time, and thus, the detected results are not timely. With the emergence of artificial

intelligence (AI) techniques since the last decade, researchers have begun to adopt multivariate

linear regression (MLR), artificial neural networks (ANN), adaptive neuro-fuzzy inference

system (ANFIS), and Fuzzy time series (FTS) model to predict water quality by exploring the

linear and non-linear relationships residing in water quality datasets. In addition, the wavelet

denoising method and intelligent algorithms are also proposed to combine with machine learning

techniques to enhance the prediction accuracy. In the following, we will review these related

work in four categories of machine learning methods.

2.1 MLR

MLR is a kind of statistical analysis method which is used to estimate the target value based

on given values collected from a set of independent variables. It is adopted to predict the water

quality because of its speed and simplicity. In [3], the MLR model is used to predict biochemical

oxygen demand (BOD) and chemical oxygen demand base on four independent variables,

temperature, pH, total suspended solid, and total suspended. The system quickly receives

relatively good result in BOD prediction with a correlation coefficient value of 0.5. MLR model

has also been used to predict the water quality index in [10] and found to be reliable in

formulating the relationship excluding the parameter chloride. However, the MLR model can

only be used to formulate linear relationship. It is likely to have a large prediction error if the

8

MLR model is used to predict non-linear relationship.

2.2 ANN

Various ANN models have been designed to predict water and wastewater discharge quality

based on previous existing datasets. A two-layer ANN model has been applied to predict the DO

concentration in the Mathura River [11], and the experimental result showed that the ANN

model worked well. In [12], various neural network types are compared in predicting water

temperatures in streams. A radial basis function neural network has also been proposed to

describe the water quality parameters in [13]. The summary of the experiment result shows the

model outperforms the linear regression model in conductivity, turbidity, and total dissolved

solids prediction. A time series prediction model, namely the autoregressive integrated moving

average, was integrated with the ANN model to improve the prediction performance. The

experimental results showed that the hybrid model provided better accuracy than ARIMA and

ANN models [14]. Additionally, a comprehensive comparison between ANN and MLR models

in biochemical oxygen demand and chemical oxygen demand prediction has been performed [3].

The experimental results show that a three-layer neural network model outperforms an MLR

model. The other comparison between ANN and MLR models in water quality index prediction

furtherly proves that the ANN model is a better option [10].

Although ANN models can effectively improve the prediction accuracy of water quality

parameters, shortcomings still exist. Especially in some scenarios where the input parameters are

ambiguous, neural networks struggle to formulate a non-linear relationship. In [15], wavelet

transformation was applied to the ANN model to improve the prediction accuracy of a variety of

ocean water quality parameters. An integration of a particle swarm optimization algorithm with

ANN models has also been investigated to improve the forecasting performance [16]. In [17],

120 data samples, collected from 2002 to 2012, are used to verify whether the integration of

fuzzy logic and ANN models can improve the water quality prediction performance. The

experimental results confirm that the proposed method works.

2.3 ANFIS

Many studies have proven that ANFIS, which can

integrate linear and non-linear relationships

hidden in the dataset, is a better option in this scenario [5]. The experimental results in [6] show

that an ANFIS model works much better than an ANN model in predicting dissolved oxygen,

even though there are only 45 data samples available. An ANFIS model with eight input

parameters is used to predict total phosphorus and total nitrogen, the experiment result based on

120 water samples shows the proposed model is reliable [18]. The ANFIS model has also been

applied to estimate the biochemical oxygen demand in the Surma River [19]. The testing results

from 36 water samples confirmed that the ANFIS model could accurately formulate the hidden

relationship and correlation analysis can improve the prediction accuracy. Two different kinds of

ANFIS model, fuzzy c-means and subtractive clustering-based was compared in [20], the

experiment result shows the ANFIS model built by fuzzy c-means provides more accurate

prediction result. In [21], the ensemble models of wavelet ANNs are found to be superior to the

best single model for forecasting chlorophyll and salinity concentrations in coastal water. An

ensemble of ANN and ANFIS is proposed in [22] to improve the prediction performance of the

ANN and ANFIS model, the test result shows there is a significant improvement in the Ensemble

ANN-ANFIS model.

According to the developer of the ANFIS model, the size of the training dataset should be no

less than the number of training parameters [23]. In the aforementioned papers, though the

ANFIS models have received higher prediction accuracy, the sizes of the training datasets are