Water Quality Classification Using Data Mining Techniques: A Case Study on Wang River in Thailand

Kittichai Northep
Department of Computer Science
Faculty of Science and Technology
Thammasat University
Pathumthani, Thailand
kittichai.n@sci.tu.ac.th

Krittakom Srijiranon*
Department of Computer Science
Faculty of Science and Technology
Thammasat University
Pathumthani, Thailand
krittakom@cs.tu.ac.th

Narissara Eiamkanitchat
Department of Computer Engineering
Faculty of Engineering
Chiang Mai University
Chiang Mai, Thailand
narisara@eng.cmu.ac.th

***Abstract*—In order to survive, the creatures need water, food, air, and residency. However, there are many water crises that influence water quality or cause water shortages. Rivers, which are important freshwater sources for human consumption, are often waste caused by their activities. This research utilizes data mining techniques to create classification models for water quality issues. The results are used in the application to alert the public about water quality problems that are aimed at changing their behavior. The data set consists of nine input features to identify dissolved oxygen in the next quarter, divided into two levels: "good" and "bad". The process of data preparation using various methods applies to the raw data before creating the classification model. This research proposes neural networks with a multilayer perceptron and k-nearest neighbor as a classifier called MLP-kNN. The results show that the proposed model is more effective when compared to various MLP algorithms. The classification rate is more than 0.95 while the F-score of both two classes is more than 0.9. Finally, the proposed model is implemented on the web application to report and prepare for water utilization planning.**

Keywords—Water quality classification; Data Mining; Neural networks; k-nearest neighbor.

Contribution—Application of data mining classifies a water quality on the Wang River, Thailand.

I. INTRODUCTION

Data mining is the process to discover knowledge from data and there are many techniques such as classification, prediction [1]. Nowadays, data mining widely used in many real-world applications such as the character classification of Bangla language [2] [3], health classification of dairy cattle [4], weather prediction in Dhaka [5], temperature prediction [6]. Low-quality water can directly and in a roundabout way influence human life and cause numerous issues. Decreasing the production of food from aquatic animals or contaminants in aquatic animals cannot be eaten resulting in food security problems. Public health problems of water-related illnesses from bacteria such as Escherichia coli (E. coli), Staphylococcus aureus causes of diarrhea [7], [8]. The Wang River is one of the major rivers in northern Thailand that encounter water quality problems. The cause of releasing wastewater from communities and industry [9]. Such problems affect the health of Thai people who consume river water for agriculture, fishery for food, transportation.

Water quality index (WQI) is used to report water quality levels for non-environmental specialists. There are several WQI that are different in each country and category such as public indices, specific consumption indices, planning indices. National Sanitation Foundation International (NSF), the certification organization from the United States, developed the National Sanitation Foundation Water Quality Index (NSF-WQI) which the one of popular WQI. NSF-WQI calculates values from nine water quality parameters which different weights for each parameter [10] [11], while the Thai government has developed the Thai Water Quality Index (TWQI) from 5 water quality parameters [12].

This research proposes the data mining process for classifying the Dissolved Oxygen (DO), one of the water quality parameters, which is the highest weight of the NSFWQI parameters and included in the TWQI parameter. The neural network with the multi-layer perceptron (MLP), demonstrated high performance in water quality prediction [13], combined with the k-nearest neighbor (k-NN) to classify data in the next quarter of the year. Eventually, the whole process is implemented on a web application to alarm people and to plan for wastewater treatment.

II. LITERATURE REVIEW AND RELATED WORK

There are many numbers of research that applied data mining processes for water quality prediction and classification of different rivers in the world.

Wavelet De-noising Technique with the Adaptive Neuro-Fuzzy Inference System (WDT-ANFIS) was proposed by Ahmed et al. [13] to predict three water quality parameters including Ammoniacal Nitrogen (AN), Suspended Solids (SS) and pH. Raw data, including twelve water quality parameters, were collected from the four data monitoring stations in the Sungai Johor River, one of the principal rivers of Johor state, Malaysia. They compare the proposed model with the three other ANN structures, including the Multilayer Perceptron (MLP-ANN), Radial Basis Function (RBF-ANN), Adaptive Neuro-fuzzy Inference System (ANFIS). They compare two types of input parameters in the data preparation process. The results show that the original input plus focus parameters from the upstream station has less error than using only the original input. Several experiments in the model creation process are

*Corresponding Author

978-1-7281-9331-1/20/$31.00 ©2020 IEEE

used to find the appropriate structure with fewer errors, such as finding activation functions for neural networks and membership functions for fuzzy logic, finding the number of nodes in each layer.

MLP-ANN was proposed by Singh et al. [14] to predict two water quality parameters including, Biochemical Oxygen Demand (BOD) and DO. Raw data, including eight water quality parameters, were collected from eight data monitoring stations in Gomti, India. Haghiabi et al. [15] reported that the support vector machine (SVM) is better performance than neural networks, including MLP-ANN and group method of data handling (GMDH) for raw data from the Tireh River, Iran. They select 8 water quality parameters as input and output parameters. They select trial and error methods with different kernel functions in order to find the appropriate kernel functions to create a SVM model. Both of these research select two model evaluation methods, Root Means Square Error (RMSE) and R-squared (R^2) to show model performance. Najafzadeh et al. [16] proposed an Evolutionary Polynomial Regression (EPR) to predict three water quality parameters BOD, Chemical Oxygen Demand (COD), DO. Raw data, including nine parameters, were collected from eight data monitoring stations in Karoun River, Iran. The proposed model is compared with RT and Gene Expression Programming (GEP). This research selects four model evaluation methods, including RMSE, mean absolute percentage of error (MAPE), correlation coefficient (*r*), Nash–Sutcliffe efficiency (NSE).

MLP-ANN is selected to predict COD values from other water quality sources, such as wastewater treatment plants proposed by Bekakari et al. [17]. Raw data, including six water quality parameters. There are several experiments to create an appropriate MLP-ANN structure, including four training algorithms, three activation functions, and a different number of hidden nodes. The hybrid model between the Regression Tree (RT) and Support Vector Regression (SVR) called RT-SVR, was proposed by Chakraborty et al. [18]. It uses to predict the pH from boiler water quality in the wastewater treatment plants. Raw data includes 5 parameters. The results of the proposed model are compared with 5 other models, such as MLP-NN, SVR, RT. Chen et al. [19] found a relation between each water quality parameter by Principal Component Analysis (PCA). There are nine water quality parameters from the 32 confined water quality monitoring samples in Jining City, China.

III. Overview of the Water Quality Parameters in Wang River

A. Study area and data monitoring

The significant rivers in the north of Thailand are the Ping River, Wang River, Yom River, and Nan River. The source of the Wang River is the Phi Pan Nam mountain range, which is in Wiang Pa Pao District, Chiang Rai Province. General information of the Wang River consists of a water route that flowing from north to south, one of the main rivers in Lampang province, and joins to the Ping River, a tributary of the Chao Phraya River which an important river of Thailand in Tak province.

The raw data in this study was courtesy of the Water Quality Management Bureau (WQMB), Pollution Control Department, Ministry of Natural Resources and Environment, Thailand. There are six water quality monitoring stations located in various points in the Wang River as shown in Fig. 1. There are five stations which compounding Thong Sawat Bridge (WA02), Sopprap Waterworks Authority (WA03), Yang Dam (WA4.1), Setu Wari Bridge (WA5.1), Ban Luk (WA06) are in Lampang province and one station, Banwangman (WA01) Bridge, in Tak province.

The raw data collected between 2001 and 2019 are divided into two periods according to the data collection period. The first phase is between 2000 and 2007. The raw data are collected between 1 and 5 times per year and differences in each water quality monitoring station. The second phase is between 2008 and 2019. The raw data are collected four times per year except for the 2009 collected six times. WQMB does not specify a date for raw data to be collected, so the period of collected data is different in each year. This study summarizes the duration that WQMB gathers raw data as follows:

- The first time is collected from February – March.
- The second time is collected from May – June.
- The third time is collected from August – September.
- The fourth time is collected from November – December.
- In 2009, WQMB collects raw data six times, including January – February, March, April, May, August, December while there are two records in 2019.

Figure 1. The six data monitoring location on the map

B. The detail of water quality parameters

From the preliminary analysis of raw data, this research selected nine relevant water quality parameters [20] as follows:

- Dissolved Oxygen (DO) is the amount of oxygen in the water.

- Fecal Coliform Bacteria (FCB) such as E. coli are bacteria found in the excretion of organisms that cause diarrhea.
- Total Coliform Bacteria (TCB), including FCB and non-fecal coliform bacteria, such as Enterobacter aerogenes (E. aeogenes), found in soil or plant excretion.
- The Potential of Hydrogen ion (pH) is the level of acidity or basicity.
- Biochemical Oxygen Demand (BOD) is the amount of oxygen consumed by microorganisms to decompose organic matter.
- Temperature (Temp) is a heat level in the water.
- Total Phosphate (TP) is the total phosphorous that causes by cycle and anthropogenic activities
- Turbidity (Tur) is the cloudiness level of water that suspended matter blocking sunlight.
- Total Solids (TS) including suspensions solids and dissolved solids

Each input parameter is different in the mixing unit, milligram per milliliters (mg/L), Nephelometric turbidity unit (NTU), the degree Celsius (°C), Most probable number per 100 milliliters (MPN/100mL). The basic statistics of the parameters are presented in Table I, including the minimum (Min), maximum (max), mean, standard deviation (SD), coefficient of variation (CV). The CV calculated by (1) is defined to dispersion around the mean of data expressed as a percentage. The high CV means a large change between each record of data. The CV is between 7.59% and 320.98%, which shows the high seasonal influences causing different anthropogenic activities. The pH and Temp are quite stable and change according to the season, while both FCB and TCB have a wider range and high dispersion of values than other parameters.

$$CV = \frac{SD}{mean} \times 100 \quad (1)$$

TABLE I. BASIC STATISTICS OF WATER QUALITY PARAMETERS IN WANG RIVER

Parameters (unit)	Min	Max	Mean	SD	CV
DO (mg/L)	2.80	11.20	6.68	1.49	22.23
BOD (mg/L)	0.05	7.40	1.43	0.79	55.45
TP (mg/L)	0.01	1.34	0.09	0.13	142.20
TS (mg/L)	54.00	1.07e+3	251.29	116.01	46.17
Tur (NTU)	0.00	999.00	79.23	117.57	148.39
Temp (°C)	21.00	34.60	28.57	2.50	8.74
pH	3.90	9.00	7.60	0.58	7.59
FCB (MPN/100mL)	20.00	1.60e+5	5.17e+3	1.66e+4	320.98
TCB (MPN/100mL)	40.00	1.60e+5	1.64e+4	3.30e+4	201.73

C. Calculation of water quality index in Thailand

In 2016, the Thai government developed a new water quality index based on historical data called the Thai Water Quality Index (TWQI) [12]. TWQI is divided into 4 periods, the information and utilization shown in Table II. The score of TWQI is between 0 and 100, calculated from the average scores of five water quality parameters, including DO, BOD, TCB, FCB, NH_3-N plus special scores.

Each water quality parameter has four meanings that share the same TWQI while using different equations to calculate the TWQI scores due to the values and units. For example, DO has six equations and some of the equations shown in (2) where DO is DO in mg/L and DO_{TWQI} is the TWQI scores while FCB has only four equations.

$$DO_{TWQI} = \begin{cases} 5 * DO + 41 \; ; \; 4.1 \leq DO \leq 6.0 \\ 12.083 * DO - 1.5 \; ; 6.1 \leq DO \leq 8.4 \end{cases} \quad (2)$$

TABLE II. THAI AIR QUALITY INDEX VALUE AND MEANING

Class	TWQI	Utilization	The measured value	
			DO	FCB
Good	71 - 100	water sports, fishery, consumption*	5.9 – 8.7	0 – 1000
Fair	61 - 70	agriculture, consumption*	4.0 – 5.8, 8.8 – 8.9	1001 – 4000
Marginal	31 – 60	industry, consumption*	2.0 – 3.9, 9.0 – 11.2	4001 – 90000
Poor	0 - 30	transportation	0.0 – 1.9, > 11.2	> 90000

*with disinfection and water quality improvement processes

IV. CONCEPT OF DATA MINING AND SYSTEM ARCHITECTURE

This research proposes the application of techniques and processes in data mining science for water quality data from the Wang River in Thailand. There are four processes including data preparation, model creation, model evaluation, and model implementation.

A. Data preparation

This process consists of several sub-processes to apply raw data before creating a model, because the raw data is usually dirty from incomplete and noisy data from many factors, such as unavailable measuring instruments or a lot of missing data in the collecting process.

First, the data cleaning processes are performed to check and eliminated incomplete data from raw data. There is a total of 446 records from six stations of WQMB, but when considered in detail, there are many missing data in some water quality parameters. For example, NO_3-N has 113 records, of which approximately 26% of the raw data or NO_2-N has 90 records, which around 20% of the raw data. This study selects only nine parameters with missing data below 5% from raw data. Finally, nine parameters with 446 records are scanned and eliminated the records that have missing data. Therefore, there are around 45 – 50 records per station.

Second, the data transformation processes to rescale data. There are high and variety of CV, different units, and a wide range of parameters. This research selects three methods to rescale the raw data, First, min-max normalization rescale data by min and max value. The true minimum and maximum values are not chosen but this research increases maximum and decreases minimum value by 10%. For example, the actual minimum and max maximum of Temp are 21.00 and 34.60 respectively. The min and max for normalization are 18.90 and 38.06, respectively. Second, the z-score rescales data by mean

and SD from. The last method is binning by the width that it divides the data into a group or called bin and each bin are an equal range of value

Finally, the data reduction with the Pearson correlation coefficient (r) is used to find the strength of the linear relation between the two parameters. The value of r is in a range between -1 and 1. The positive correlation occurs where $r > 0$ and the negative correlation occurs where $r < 0$. The $|r|$ is near 1 showing strong relationships while the $|r|$ is near 0 showing a weak relation.

B. Model creation

Multi-layer Perceptron (MLP) is the structure of an artificial neural network model for supervised learning. It consists of the input layer, hidden layer, an output layer, that each layer has the number of nodes depending on the data. The activation function transfers the summation of every input and their weights in a node (a) to the new value. Three functions are selected: logistic sigmoid (Sig) with ranges (0, 1), hyperbolic tan ($Tanh$) with ranges (-1, 1), and rectified linear unit ($ReLU$) with ranges (0, infinity) showing on (3).

$$f'_{ReLU}(a) = \max(0, a) \tag{3}$$

The backpropagation algorithm used in the learning process to update weights from the errors after learning. There are two weight optimization methods, a stochastic gradient-based method proposed by [21] called Adam, and a Limited-memory Broyden–Fletcher–Goldfarb–Shanno method proposed by [22] called L-BFGS.

The k-nearest neighbor (k-NN) classifier is a simple and acceptable performance. The different distance functions of k-NN affect the accuracy in classification depending on the type of data [23]. It classifies class by finding the distance between unknown data and k nearest data. The variable k is the number of data, then data are identified by voting. There are several distance metrics and Minkowski distance ($D(X,Y)$) from (4) is selected where n is the number of features, p is order, X and $Y$are record of data.

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \tag{4}$$

C. Model evaluation and model implementation

The created model must check the performance between actual value (y) and model value (y'). The R-squared (R^2) represented by (5) is metrics for the prediction model, where $\overline{y}$ is the average value of actual value.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y-y')^2}{\sum_{i=1}^{n}(y-\overline{y})^2} \tag{5}$$

For the classification model, a confusion matrix is a table showing the number of classifying data between actual class and prediction class. In the two-classes model, the confusion matrix has two rows and columns called true positive (TP), false negative (FN), true negative (TN), false positive (FP). It can calculate other metrics, including accuracy, precision, recall. F-score is represented by (6) is a popular metric to describe classification performance.

$$F - score = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \tag{6}$$

Finally, the web application is developed and the best model for each station is implemented.

V. METHODOLOGY

This section is an experiment in each process of the data mining techniques to find an approximate model. The first subsection shows the preliminary results of both classification and prediction. The next subsection identifies the type of input data and the objective of the model. Next, many experiments have been done to find the best structure of MLP. The k-NN classifier is selected to compare with other algorithms. Finally, the web application is developed.

A. The preliminary experiment

The objective of this experiment is to be used as a basic guideline in the data mining process for water quality data. Data sets are created to DO prediction and classification in the next quarter. The min-max normalization method of data transformation is chosen to rescale the data set. The range of data set after rescaling depends on the activation function of BPNN, such as between 0 and 1 for the sigmoid activation function. The structure of BPNN is created from a trial and error method. The sigmoid activation functions and two hidden layers with fixed ten hidden nodes are selected. The 5-fold cross-validation is divided data set into training and testing groups.

The output of BPNN is compared with the actual value to calculate R^2. The output of BPNN is converted to four classes from Table II for classification performance calculations. The results show in Fig. 2, where the z-score has an R^2 value higher than min-max normalization on five stations except WA03, but all stations have an extremely low R^2 value. The highest value of R^2 is only 0.2959 at WA01, showing very low model performance.

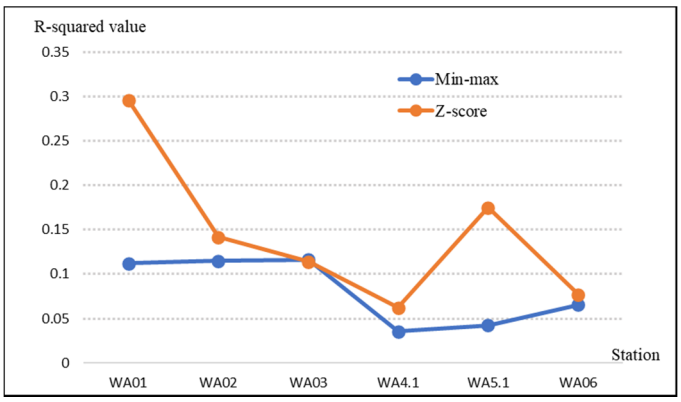

Figure 2. The R-squared value of DO prediction on six stations

There is also a major problem in classification due to unbalancing, the number of classes, and causing some stations to miss classification. The major class is around 60 – 70% always in class "Good" while minor classes are around 0 – 45%. The result of the classification shows that the model cannot classify data in minor classes.

B. *Type of input feature and output data*

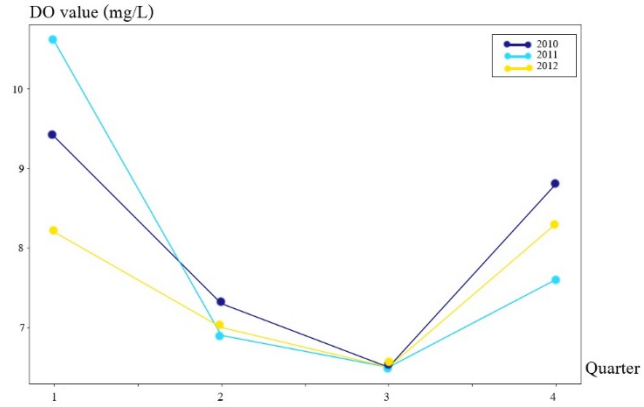

Figure 3. The 3-year trend of DO of station WA01 dividing by quarter

From the preliminary experiment, input data are selected from the current quarter to predict the output of the next quarter. For example, the input data is the first quarter of 2019, and output data is the second quarter of 2019. Due to low performance, the past year of data is selected as input data while output data is the next quarter of the current year. For example, the input data is the first quarter of 2018, and output data is the first quarter of 2019. The model was created from past year data with the same structure as the previous experiments. The model performance of past year data is more efficient and, Fig. 3 shows the trend of data divided by the quarter between 2010 and 2012 of station WA01. The first quarter has high DO values. There is a sharp drop in the second quarter and slightly decreases in the third quarter respectively, while the last quarter is rapidly increasing.

Reporting the prediction value of both two parameters to non-technician is difficult to understand and has a very low performance. From the previous section, there are four classes of water quality and Table III shows the number of data in each class of the DO. Most of the four-level problems are mostly no data in the "Poor" class, while there is little data in class "Fair" and "Marginal". Therefore, this research combines class "Poor", "Fair", and "Marginal" to class "Bad". Finally, there is only two-class "Good" and "Bad".

TABLE III. THE NUMBER OF DATA CLASS DIVIDING BY STATIONS AND CLASSES

Station	DO				The proportion of class (Good : Bad)
	1	*2*	*3*	*4*	
WA01	29	23	0	0	56 : 44
WA02	39	5	7	0	76 : 24
WA03	39	4	8	2	74 : 26
WA4.1	33	12	3	0	69 : 31
WA5.1	37	11	6	0	68. : 31
WA06	44	5	2	0	86 : 14

C. *Correlation analysis and feature reduction*

Due to poor model performance of a preliminary experiment. The descriptive analysis of the data set is conducted. May be due to a feature that does not provide information and only a small amount of data is recorded in each station. Feature reduction method with r is selected to reduce unnecessary features. First, the raw data are calculated r between nine input features and two outputs. The results show that the r values of DO are between -0.0985 and 0.3850, while the highest correlation is TUR and the lowest correlation is TP. The r values of FCB are in a range between -0.1731. and 0.2608 while the highest correlation is DO and the lowest correlation is TCB. In conclusion, the correlation is very low in every feature.

Next, the Binning method is applied to raw data before finding r. Each feature is divided into a bin by equal width which the number of bins between 2 and 5. The data smoothing process is applied to each bin by converted to bin means. Finally, the Pearson correlation coefficient is selected to find correlation from smoothing data between input features and outputs. For feature selection, the r value greater than 0.8 between the input feature and output is selected to create the classification model. It is divided correlation results into three groups, combining both positive and negative as follows:

- a high correlation in every station: Temp, pH, DO
- a high correlation in some stations: Tur, BOD, TCB, TS
- a low correlation in every station: FCB, TP

Each station selects the difference in the input features depending on its r value with only a high correlation. For example, station WA01 selects Temp, pH, DO, BOD, and TS, while station WA4.1 selects Temp, pH, DO, and TCB.

D. *The experiment for appropriate MLP structure*

The only feature from raw data that more r than a threshold is selected, then two data transformation method with min-max normalization and z-score are applied and compared to the non-data transformation. To create an appropriate MLP model, there are two types of MLPs. First, one MLP structure with a hidden layer and two output nodes, including one node for class "Good" and the other one for class "Bad". This type is called "MLP-1". The second is two MLP structures with hidden layers and one output node called "MLP-2". The objective of the first structure is to recognize class "Good" to 1 and the second structure is the objective for recognition class "Bad" to 1. The maximum method of output between the two output values is selected to identify the class.

The MLP-1 was created with several hidden layers, that vary between 1 and 5. Three activation functions including ReLU, Sigmoid, Tanh were used in experiments. The number of nodes in each hidden layer is fixed by $2n + 1$, where n is the number of input nodes. The data set is divided into training and testing with 5-fold cross-validation. Accuracy and F-score are used as indicators for comparing the model performance.

First, only station WA01 and Adam weight optimization are selected because there are many numbers of class "Bad" in WA01. The non-data transformation is very low performance while min-max normalization outperforms other methods. The sigmoid activation function is less performance than the other two activation functions, while ReLU and Tanh are acceptable performance. The one and two hidden layers of MLP are lower

performance models, while the five hidden layers of MLP is the best performing. Due to the page limitations, only important experiment results are presented.

In the next step, the weight optimization L-BFGS is selected to compare with Adam. The MLP with five hidden layers, two different weight optimization methods, and two activation functions are selected. The model performance of each fold of station WA01 shows in Table IV for the classification rate and Table V for F-score focusing on class "Good" as the positive class. The results show that L-BFGS is higher model performance than Adam in most fold excluding fold 5 on both two-performance metrics. The Tanh activation function is slightly better than ReLU.

TABLE IV. A CLASSIFICATION RATE OF STATION WA01 WITH DIFFERENT STRUCTURE OF MLP

WO*	AF**	A classification rate of a fold					
		1	*2*	*3*	*4*	*5*	*Avg.*
Adam	ReLU	0.70	0.56	0.78	0.50	0.67	0.64
	Tanh	0.80	0.56	0.56	0.63	0.63	0.63
L-BFGS	ReLU	0.90	0.67	0.89	0.88	0.63	0.79
	Tanh	0.80	0.78	1.00	0.88	0.63	0.82

*WO: Weight optimization method and **AF: Activation function

TABLE V. F-SCORE OF STATION WA01 WITH DIFFERENT STRUCTURE OF MLP

WO*	AF**	F-score with class "Good" of a fold					
		1	*2*	*3*	*4*	*5*	*Avg.*
Adam	ReLU	0.80	0.60	0.86	0.60	0.77	0.73
	Tanh	0.83	0.60	0.71	0.73	0.67	0.71
L-BFGS	ReLU	0.92	0.73	0.92	0.91	0.73	0.84
	Tanh	0.83	0.80	1.00	0.91	0.73	0.85

*WO: Weight optimization method and **AF: Activation function

The five other stations are created MLP model with the L-BFGS weight optimization method and compare the model performance between two activation functions. Due to the small amount of data in the minor classes of three stations, it cannot calculate the F-score of class "Bad" as a positive class. As shown in Table VI, the experimental results of some fold of WA02, WA03, and WA06 stations are no value as "-", so the three other stations are selected. The model performance of three selected stations between two activation functions shown that the Tanh activation function is slightly better than ReLU.

The MLP-2 model with the same structure of the MLP-1, which uses the Tanh activation function, L-BFGS weight optimization method, and five hidden layers are created to classify three selected stations. Table VII shows classification rates compared between MLP-1 and MLP-2. The result of the average value of 5-fold cross-validation shows that MPL-1 outperforms the MPL-2 in station WA01 and WA5.1. The other results, which not showing in table, include MLP-2 outperforming F-score of class "Bad" only in station WA4.1, MLP-1 outperforming F-score of class "Good" in station WA01 and WA5.1

TABLE VI. F-SCORE OF CLASS "BAD" ON THREE NONE-SELECTED STATION

Station	AF	F-score with class "Bad" of a fold				
		1	*2*	*3*	*4*	*5*
WA02	ReLU	-	0.44	0.5	-	0.5
	Tanh	0.4	0.5	-	-	-
WA03	ReLU	0.5	0.75	0.5	-	0.4
	Tanh	0.67	0.86	0.33	-	0.4
WA06	ReLU	0.5	-	-	0.67	0.86
	Tanh	-	-	1	-	-

TABLE VII. A CLASSIFICATION RATE OF THE SELECTED STATION WITH TWO MLP STRUCTURES

Station	MLP structure	
	MLP-1	*MLP-2*
WA01	0.82	0.64
WA4.1	0.44	0.60
WA5.1	0.77	0.64

E. The results of classification with MLP-kNN

Both MLP models use k-NN as a classifier instead of the maximum method. The Minkowski distance is chosen with different numbers of p and k, with ranges 3-10 and 3-5 respectively. The MLP-1 with k-NN as classifier outperforms than MLP-1 with a maximum method as classifier only on results from station WA4.1. The MLP-2 with k-NN as a classifier, also known as MLP-kNN, is a high classification performance. There are different k-NN structures depending on the station that the best structure is shown in Table VIII. The performance is stable when changing the p value, while the lower k value is the lower performance. The classification rate is very high in WA01 and WA5.1 stations that are greater than 0.95, while station WA4.1 is low value with 0.6861. F-scores are reported in two types, when considering the good and bad class as the positive class. There are more than 0.9 at station WA01 and WA5.1, but there is less performance at WA4.1 station.

TABLE VIII. THE AVERAGE CLASSIFICATION PERFORMANCE OF MLP-KNN

Station	MLP-kNN		F-score		Classification rate
	k	*p*	*Good*	*Bad*	
WA01	9	3	0.9778	0.9714	0.9750
WA4.1	10	5	0.7833	0.8044	0.7563
WA5.1	10	3	0.9667	0.9333	0.9556

F. *The model performance comparison results*

Support vector machine (SVM) with different kernel functions, including linear, sigmoid, Radial Basis Function (RBF), and decision tree (DT) with different loss functions including the Gini coefficient, cross-entropy are selected to the created classification model. The best structure comparison of SVM with MLP-1 and MLP-kNN is shown in Fig. 4, Fig. 5, and Fig. 6 for stations WA01, WA4.1, and WA5.1 respectively, with the average value of 5-fold cross-validation. Only station WA4.1 has MLP-1* with k-NN as a classifier instead of MLP-1.

The result of SVM shows that station WA01 is not able to calculate the F-score of class "Bad" with all kernel functions, station WA4.1, and WA5.1 are the best performance with RBF kernel functions, but there is lower performance than other algorithms. The results of DT show that the loss function with cross-entropy has higher model performance than the Gini coefficient, station WA5.1 cannot calculate F-score of class "Bad", station WA4.1 is a slightly lower performance metric than the proposed model, station WA01 is poorly performance. Due to a few of the numbers of data in class "Bad", every model except the proposed model is low F-score value of class "Bad". The results of the proposed model show that it is the highest performance than other algorithms in every performance metric and stations, but station WA4.1 has lower performance than the others because there is a very low recall of class "Good" in three folds. The proposed model has high efficiency for classifying class "Bad".

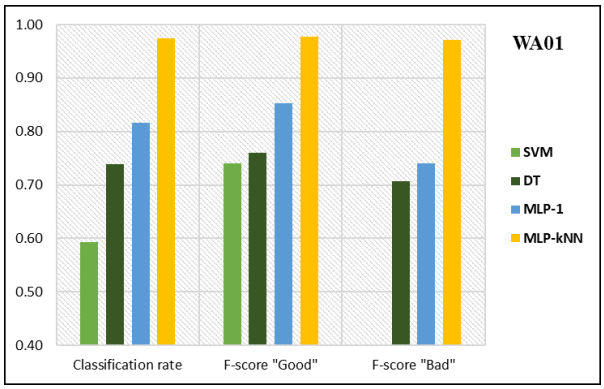

Figure 4. Comparision result between three algorithms of station WA01

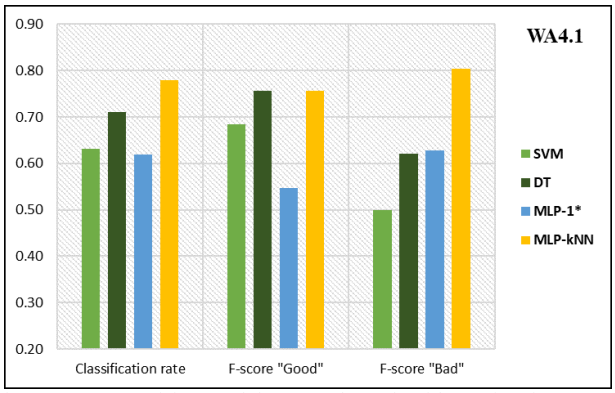

Figure 5. Comparision result between three algorithms of station WA4.1

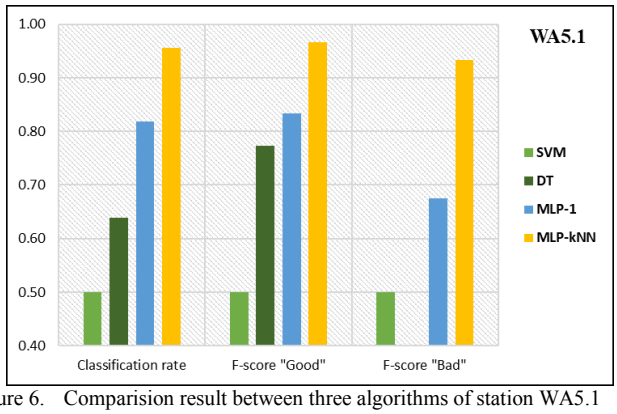

Figure 6. Comparision result between three algorithms of station WA5.1

G. *Proposed system architecture*

The proposed system architecture shows in Fig. 7. First, the raw data consists of nine parameters related to the water quality, from six monitoring stations. The data preparation process eliminates a record of data that has a missing value. Binning method rescales raw data and feature selection with *r* is selected high correlated between input features and DO. The selected features are different in each station. For example, five features for station WA01, six features for station WA5.1. In the model creation process and model evaluation, many experiments were created to find the approximate structure of MLP-1 and MLP-kNN by two model performance metrics accuracy and F-score. The best structure is MLP-kNN, which different *k* values depending on the station are implemented to the web application.

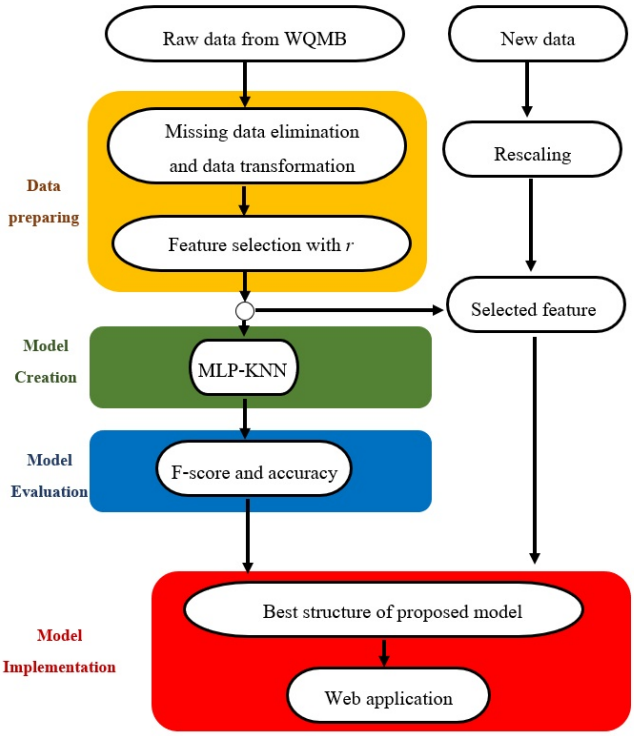

Figure 7. The proposed system architecture

The web application is developed by React, JavaScript library from Facebook, and MongoDB, NoSQL database. The

proposed MLP-kNN is implemented on the backend, while Fig. 8 shows the front end, which includes the location of six stations with current water quality data, predicted class of DO in next quarter, the historical statistic data.

VI. CONCLUSION

This research applies the data mining techniques to classify the dissolved oxygen of the next quarter from historical data. Due to a lack of the number of data, feature reduction with the Pearson correlation coefficient is used to eliminated uncorrelated features. Min-max normalization is applied to the data set in the data pre-processing step. The MLP-kNN is a proposed model created from numerous experiments that shows good classification performance. F-score of class "Bad" is a focus metric performance during the created model because it is a minor class that affecting public health. Due to the lack number of a minor class, only three stations are selected in this research. The web application is developed and implemented as a prototype. The objective is to identify water quality and planning for the future.

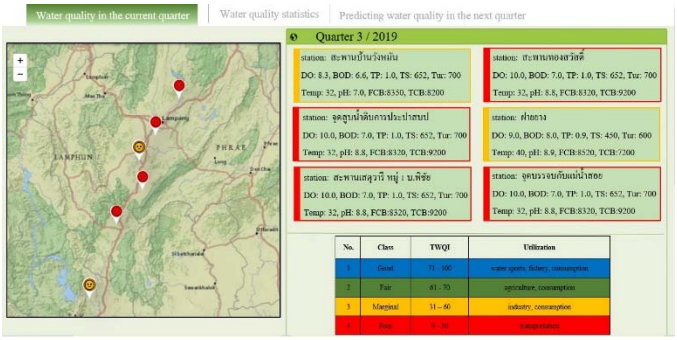

Figure 8. The user interface of web application

ACKNOWLEDGMENT

The water quality data in this work was courtesy from the Pollution Control Department, Ministry of Natural Resources and Environment, Thailand. In addition, this work was supported in part by Thammasat University.

REFERENCES

[1] Y. Zhao, *R and Data Mining: examples and Case Studies*. Academic Press: San Diego, 2013.

[2] C. Saha, R. H. Faisal and M. M. Rahman, "Bangla Handwritten Basic Character Recognition Using Deep Convolutional Neural Network," *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Spokane, WA, USA, 2019, pp. 190-195.

[3] R. R. Chowdhury, M. S. Hossain, R. ul Islam, K. Andersson and S. Hossain, "Bangla Handwritten Character Recognition using Convolutional Neural Network with Data Augmentation," *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Spokane, WA, USA, 2019, pp. 318-323.

[4] A. Pimpa, N. Eiamkanitchat, C. Phatsara, and T. Moonmanee, "Decision Support System for Dairy Cattle Management Using Computational Intelligence Technique," in *Proc . The 7th International Conference on Computer and Communications Management (ICCCM)*, Bangkok, Thailand, 2019, pp. 181-185.

[5] M. Rahman, A. H. M. S. Islam, S. Y. M. Nadvi and R. M. Rahman, "Comparative study of ANFIS and ARIMA model for weather forecasting in Dhaka," *2013 International Conference on Informatics, Electronics and Vision (ICIEV)*, Dhaka, 2013, pp. 1-6.

[6] Tushar Kanti Routh *et al*., "Artificial neural network based temperature prediction and its impact on solar cell," *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*, Dhaka, 2012, pp. 897-902.

[7] "Quality Unknown: The Invisible Water Crisis," *World Bank*. [Online]. Available: https://www.worldbank.org/en/news/feature/2019/08/20/quality-unknown. [Accessed: 20-Dec-2019].

[8] P. P. T. V. Online, "The danger of wastewater," *pptvhd36.com*, 27-Jun-2019. [Online]. Available: https://www.pptvhd36.com/news. [Accessed: 20-Dec-2019].

[9] Thairath Online, "Poor water quality in Wang River. Don't catch fish for food," *www.thairath.co.th*, 10-Apr-2016. [Online]. Available: https://www.thairath.co.th/news/local/603995. [Accessed: 20-Dec-2019].

[10] T. Poonam, B. Tanushree, and C. Sukalyan. "Water quality indices-important tools for water quality assessment: a review," *International Journal of Advances in Chemistry*, vol. 1, no. 1, pp. 15-28, 2013.

[11] A. Said, D. K. Stevens, and G. Sehlke, "An Innovative Index for Evaluating Water Quality in Streams," *Environmental Management*, vol. 34, no. 3, pp. 406–414, 2004.

[12] Water Quality Management Bureau, Pollution Control Department, "New WQI calculator," *Inland water quality information system*. [Online]. Available: http://iwis.pcd.go.th/index.php?method=publications&etc=1574757413456. [Accessed: 26-Nov-2019].

[13] A. N. Ahmed, F. B. Othman, H. A. Afan, R. K. Ibrahim, C. M. Fai, M. S. Hossain, M. Ehteram, and A. Elshafie, "Machine learning methods for better water quality prediction," *Journal of Hydrology*, vol. 578, p. 124084, 2019.

[14] K. P. Singh, A. Basant, A. Malik, and G. Jain, "Artificial neural network modeling of the river water quality—A case study," *Ecological Modelling*, vol. 220, no. 6, pp. 888–895, 2009.

[15] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, "Water quality prediction using machine learning methods," *Water Quality Research Journal*, vol. 53, no. 1, pp. 3–13, 2018.

[16] M. Najafzadeh, A. Ghaemi, and S. Emamgholizadeh, "Prediction of water quality parameters using evolutionary computing-based formulations," *International Journal of Environmental Science and Technology*, vol. 16, no. 10, pp. 6377–6396, Oct. 2018.

[17] N. Bekkari, and A. Zeddouri, "Using artificial neural network for predicting and controlling the effluent chemical oxygen demand in wastewater treatment plant," *Management of Environmental Quality: An International Journal*, vol. 30, no. 3, pp. 593–608, 2019.

[18] T. Chakraborty, A. K. Chakraborty, and Z. Mansoor, "A hybrid regression model for water quality prediction," *Opsearch*, vol. 56, no. 4, pp. 1167–1178, Jun. 2019.

[19] Y. Chen and L. Shu, "Confined water quality evaluation of cone of depression in jining based on principle component analysis method," *2011 International Symposium on Water Resource and Environmental Protection*, 2011.

[20] I. Ichwana, S. Syahrul, and W. Nelly, "Water Quality Index by Using National Sanitation Foundation-Water Quality Index (NSF-WQI) Method at Krueng Tamiang Aceh," *Proceeding of the First International Conference on Technology, Innovation and Society*, 2016.

[21] D. Kingma, and J. Ba, "Adam: A Method for Stochastic Optimization," *The 3rd International Conference on Learning Representations (ICLR)*, San Diego, USA, 2015, pp. 1-15.

[22] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[23] L.-Y. Hu, M.-W. Huang, S.-W. Ke, and C.-F. Tsai, "The distance function effect on k-nearest neighbor classification for medical datasets," *SpringerPlus*, vol. 5, no. 1, Sep. 2016.