Improving the Water Quality Classification Model for Various Farms Using Features Based on Artificial Neural Network

Sumitra Nuanmeesri [1], Lap Poomhiran [2], Preedawon Kadmateekarun [1], Shutchapol Chopvitayakun [1]

[1] Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand
[2] Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

***Abstract* – Measuring and classifying the water quality is necessary to manage the appropriate water quality for various farms near the coast or affected by seawater. This research aimed to improve the water quality classification model for various farms using Multi-Layer Perceptron Neural Network-based multi-class Support Vector Machine. It also implements the Random Forest Feature Importance Selection to increase model accuracy. The class reduction technique decreases the probability of co-occurrence classes for various farms in overlapping water ecosystems. The result has shown that the dataset that applied the class reduction helped increase the model's efficiency more than the feature selection technique. The models that applied the multi-class Support Vector Machine classifier are more accurate than the Softmax activation function classifier. The findings indicate that the model using Multi-Layer Perceptron Neural Network-based One-versus-One Support Vector Machine combined with the Random Forest Feature Importance Selection and the class reduction has the highest efficiency and improves the water quality classification model in various farms.**

DOI: 10.18421/TEM124-25
https://doi.org/10.18421/TEM124-25

Corresponding author: Sumitra Nuanmeesri,
Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand
Email: sumitra.nu@ssru.ac.th

Received: 17 July 2023.
Revised: 11 September 2023.
Accepted: 19 September 2023.
Published: 27 November 2023.

***Keywords* – Artificial neural network, classification, feature selection, support vector machine, water quality.**

1. Introduction

Farming in areas close to the sea that depend on water as the main factor often has problems with water quality. The cause may be the problem of dumping waste and releasing wastewater from homes and industrial factories into canals or rivers and flowing into the sea. It negatively affects water quality in a chain to aquaculture in cages, salt fields, and crop farms directly. Water quality below the optimum threshold will cause farm production rate and quality to decrease. For example, some fruit trees produce fewer and smaller fruits than before. Further, the fruit color will fade, and the taste will deteriorate. Worst is the loss of marine life in cages or the death of cultivated plants. Improper water quality for salt farms, fish farms, and crop farms may be caused by acidity-alkalinity, the potential of hydrogen (pH), salinity, water temperature (WT), dissolved oxygen (DO), or some contaminant presences, such as biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammonia (NH3), lead (Pb), copper (Cu), cadmium (Cd), mercury (Hg), and arsenic (As) that are affected to turbidity and electrical conductivity (EC).

Global warming is another factor that causes frequent climate change. It results in higher water temperatures and ocean acidification [1], [2], leading to sea level and salinity changes over time. A 2.0 degrees Celsius rise will result in water shortages [3], leading to a decrease in the amount of water that will increase the density of fish farming. In fish cages, waste products such as ammonia, nitrate, and carbon dioxide accumulate excreted from fish [4]. It also affects the growth and physiology of fish [5]. For example, the amount of DO in the water decreases, causing the fish to use much energy to exchange the air and causing the fish to grow slowly [6].

Further, weather fluctuations have caused more frequent monsoons and floods. Flooding water will reduce the salinity [7]. These floods or drainage will contain mud or sewage [8], making the fish knock out of the water or die [9]. Furthermore, increasing atmospheric carbon dioxide is dissolved in seawater. It was causing the water to have a much higher acidity [10]. Acidity negatively affects the calcium balance in the fish [11]. The salinity of the water is not balanced with minerals and the fish's functioning system affects the feed conversion ratio (FCR) [12]. Furthermore, as seawater infiltrates rivers and canals used for agriculture and gardening, it results in an elevation of salinity levels. This elevated salinity adversely affects horticultural crops, leading to their dehydration and a subsequent failure to produce fruit [13]. Besides, the monsoon situation and changes in rainfall directly affect salt farming, while salinity level intensity decreases.

In Thailand, Samut Songkhram Province is about 100 kilometers Southwest of Bangkok. It is at the mouth of the river and the sea coast. This province is influenced by rivers and seas, forming three water ecosystems: freshwater, brackish water, and saltwater. Saltwater ecosystems are influenced by seawater. Freshwater has a salinity of less than 0.5 parts per thousand (ppt); some systems define a salinity as less than 1.0 ppt, brackish water has a salinity between 0.5 ppt and 30.0 ppt, while saltwater or seawater has a salinity of more than 30.0 ppt [14]. Brackish water ecosystems are aquatic environments that contain freshwater and seawater that rises from the moon's gravitational pull, which affects the tide and ebb levels. It is typically found in areas where freshwater sources, such as rivers or streams, meet and mix with seawater bodies, such as oceans, seas, or estuaries. This mixing of waters creates a unique and dynamic habitat with characteristics of both freshwater and seawater ecosystems. Freshwater ecosystems are derived from seasonal rainfall and irrigation systems. These ecosystems cause constant water displacement, and the changing water behavior of the natural system each month of the year contributes to the abundance of biodiversity and natural resources.

Three water ecosystems in Samut Songkhram Province can be divided into lower, middle, and upper areas. The lower area is a saltwater ecosystem extending from the coastline to the inland for about three kilometers. This area is adjacent to the Gulf of Thailand and therefore is saline. Most of the villagers in the area are engaged in salt farming, mackerel fishery, shrimp farming, and fish farming. The middle area is a brackish water ecosystem. There is an area about three kilometers from the lower area. This area is supported by saltwater in the dry season. Most of the villagers in the area make a living from coconut plantations to produce sugar and fresh coconut juice. The upper area is a freshwater ecosystem. It is an area above the middle area, which is in the area of freshwater from the Mae Klong River. Villagers in the area are mainly engaged in fruit and vegetable gardening. However, farmers or villagers in overlapping water ecosystems and different farming may experience different positive and negative impacts when sharing water resources. For example, salt farming requires water with a salinity between 27-30 ppt [15], while Asian seabass (*Lates Calcarifer*) fish farms require saltwater with a salinity of 20 ppt to help these fish thrive [16], while salinity is not necessary for crop farming.

Nowadays, communication technology and the internet have significantly progressed when combined with hardware or small electronic devices. It can be controlled or managed over a global network anywhere over time. This technology is known as the Internet of Things (IoT). It allows developers or users to program and control devices or sensors via an intranet or internet connection. Most water quality collections rely on the IoT, which consists of probes and sensors that take water quality readings and return them to the processor to upload the data to a central database server or cloud server. This makes collecting various types of water quality data possible quickly and automatically. In addition, these water quality data must be analyzed using some techniques, such as machine learning or neural network processes, to provide a model that accurately classifies water quality. However, there is a wide range of water quality attributes. Some attributes may have few effects on the model's classification or may reduce its accuracy performance. Thus, feature selection is a guideline for filtering the appropriate attributes for further training of the model.

Therefore, water quality classification is essential to farmers. Suppose they have tools or models that can help alert or classify water quality accurately. In that case, it will help to prevent and reduce or alleviate the severity of the impact of water quality that is unsuitable for the various farms quickly and timely. Most studies on water quality develop models using either technique and compare the results. In addition, most of these works are only interested in water sources within the same context and affect only one activity. Besides, research on machine learning of water quality prediction is generally lacking for different farms with overlapping water ecosystems. This research presents a water quality classification model for various farms using a combination of multi-layer perceptron neural network (MLPNN) and multi-class support vector machine (MCSVM) methods, including feature selection and class reduction in the case of overlapping water sources between salt farms, fish farms, and crop farms.

This paper is organized as follows. Section 2 introduces the general concept of machine learning and related works based on feature selection in the water quality dataset. Next, section 3 describes the empirical methodology proposed. Then, section 4 presents the result. Finally, section 5 concludes and findings, respectively.

2. Literature Reviews

The problem of global warming and pollution caused by human actions affect the water condition making it not suitable for farming. These problems resulted in farmers requiring some tools to guide water quality prediction for planning their farming practices, managing water sources, and maintaining the water quality. Most popular methods rely on water quality parameters for evaluating water quality accurately. Some studies have developed IoT devices for collecting and real-time water quality monitoring [17], [18], [19]. However, not all parameters or features affect the accuracy of water quality classification. The optimal feature is critical for the model's classification performance, including feature extraction [20] and feature selection [21]. Nevertheless, the minimum feature selection has not been fixed in Samut Songkhram Province. Most studies mainly focus on reducing the number of features and observing the model's accuracy. The model will use a shorter processing time with fewer features if the precision is constant. Selecting some essential features will provide a more accurate classification than selecting all [22]. Feature selection is an essential process with the aim of reducing redundant or unimportant input features that can degrade model's efficiency in machine learning [23], [24], [25]. There are three main methods for feature selection: filter, wrapper, and embedded. Each type of feature selection method has its advantages and disadvantages. The choice of method depends on factors such as the dataset size, the dimensionality of the features, the computational resources available, and the specific goals of the analysis or machine learning task. Some studies have found that embedded methods produce better results [26], especially when using Random Forest Feature Importance Selection (RFFIS) which was developed by applied RF-based feature importance (FI).

There are several studies on water quality classification with machine learning, such as Decision Tree (DT) [18], Random Forest (RF) [25], Naïve Bayes (NB) [27], K-Nearest Neighbors (KNN) [28], Linear Regression (LR) [29], Support Vector Machine (SVM) [30], and the neural network [19]. In addition, water quality classification models were also compared between different machine learning models.

Some research found that the model using Multi-Layer Perceptron (MLP) has higher accuracy than LR, RF, and SVM [29], [31], [32]. Most models use only one machine learning classifier and compare the performance to get the best model, which includes the feature selection method. However, few studies have combined two or more machine learning within one model for classifying water quality. In particular, high-performance MLP models are combined with machine learning, such as SVM, to classify water quality further. Indeed, this study proposes a method to develop a feature-based model and use MLPNN combined with MCSVM to classify water quality for various farms, including salt farms, fish farms, and crop farms.

3. Material and Methods

Improving the water quality classification model for various farms in this research consists of five main processes: 1) data collection and preprocessing, 2) feature selection, 3) class reduction, 4) model development, and 5) model evaluation. The proposed research framework is shown in Figure 1 with the following details.

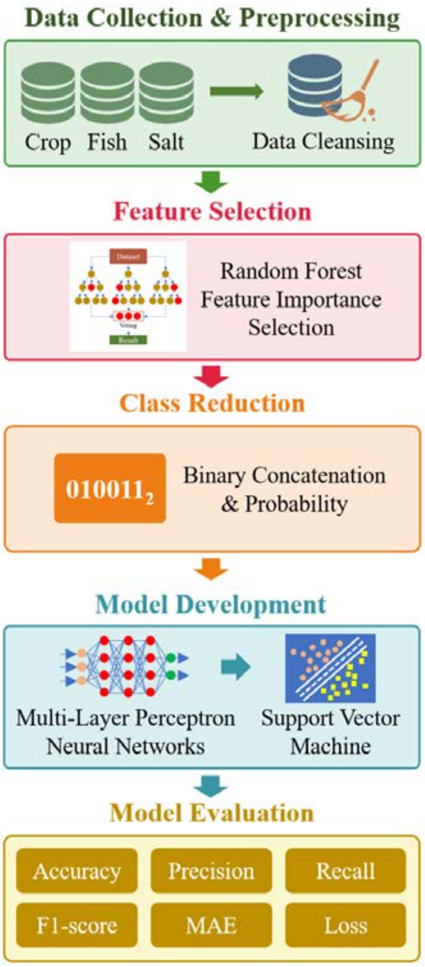


Figure 1. The proposed research framework

3.1. Data Collection and Preprocessing

Water quality data is collected by IoT devices that automatically upload data from sensors that measure water quality to a database server. These water quality parameters include salinity, pH, WT, DO, BOD, COD, NH3, Pb, Cu, Cd, Hg, As, EC, turbidity, and water flow rate, which are collected in the river and canal in Samut Songkhram Province, Thailand, between February to May 2023. The collected data undergoes a data cleansing process to remove incomplete entries such as null or missing values. There are 2,015 records after completing the data cleansing process. Following, the dataset with water quality was assigned classes for the three types of farms: crop, fish, and salt farms. Each type of farm contains three classes indicating water quality severity level: normal, caution, and critical. There are nine classes (C1 to S3) a labeled for the dataset, as shown in Table 1.

Table 1. The water quality class label for various farms

Severity	Crop farm	Fish farm	Salt farm
Normal	C1	F1	S1
Caution	C2	F2	S2
Critical	C3	F3	S3

3.2. Feature Selection

Sometimes a dataset with many attributes or features can affect the learning efficiency of the model. The dataset in this work will be processed with feature selection to find the attributes or features that significantly affect the model for use in learning the model. The RFFIS was applied in this work based on the RF and FI techniques as follows.

3.2.1. Train the Random Forest Model

The RF is an ensemble learning approach that uses several decision trees (DT) and votes the result from each tree by applying bootstrap sampling, as shown in Figure 2.

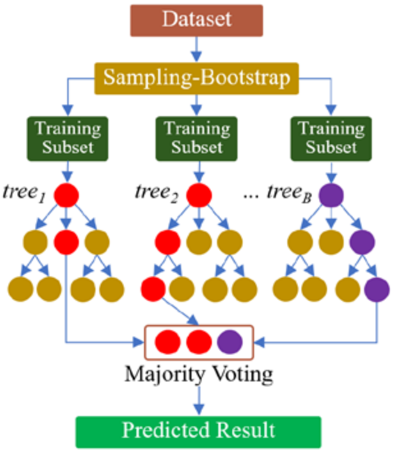


Figure 2. The Random Forest with bootstrap aggregating structure

The dataset was divided into 80% for the training set and 20% for the test set, with the random state being 42. The bootstrap re-sampling process was applied to randomize the sampling of the training subset from the dataset. The structure of the RF model was designed based on the bagging ensemble parallel processing (bootstrap sampling and aggregation). The number of estimators was set to 100 decision trees, and the maximum deep of each decision tree was set to 10. The final prediction for the input data point is determined by a majority vote among all the decision trees that vote for a class label. The class label that receives the most votes becomes the ensemble's prediction of the RF model. However, the RF model regression was applied to measure the impurity variance for feature selection in this research using the Mean Absolute Error (MAE), as in (1) [26].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{1}$$

where n is the number of output data, y_i is observed of an instance i-th, and $\hat{y}_i$ is the mean value of the observed instances.

3.2.2. Feature Importance Calculation

The feature importance was calculated by comparing Gini importance (Mean Decrease Impurity) and permutation importance (Mean Decrease Accuracy) which range between 0 and 1. The mean and standard deviation (SD) were computed in each sub-decision tree for the mean decrease. For Gini importance, the features for internal nodes were selected with Gini impurity on an average gain of the Mean Decrease Impurity (MDI). The Gini index can be calculated in (2) [33] for two classes and (3) [34] for multi-class, respectively.

$$Gini\ index = p_1(1-p_1) + p_2(1-p_2) \tag{2}$$

where p_1 is the probability of class one and p_1 is the probability of class two.

$$Gini\ index = \sum_{i=1}^{C}(p_i)^2 \tag{3}$$

where C is the total class number and p_i is the probability belonging to i-th class.

The Gini impurity can be formulated in (4) [34, 35].

$$Gini\ impurity = 1 - \sum_{i=1}^{C}(p_i)^2 \tag{4}$$

The Mean Decrease Accuracy (MDA) in impurity for permutation importance was calculated based on the permuted out-of-bag (OOB) within each tree with high cardinality features.

The permutation feature importance (PFI) with Monte Carlo for the feature vector x_j can be formulae in (5), (6), and (7) [35]

$$L^{(i)} = L(y^{(i)}, f(x^{(i)})) \tag{5}$$

$$\tilde{L}_m^{(i)} = L(y^{(i)}, f(\tilde{x}_j^{(i)}, x_{-j}^{(i)})) \tag{6}$$

$$PFI_j = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{M}\sum_{m=1}^{M}(\tilde{L}_m^{(i)} - L^{(i)})\right) \tag{7}$$

where n is the number of samples, j is the random feature, $L^{(i)}$ is the loss of prediction on the instance i-th, $\tilde{L}_m^{(i)}$ is the loss of prediction of instance m-th sample on instance i-th, and M is the number of times for estimation.

Each decision tree's feature importance was calculated and then ranked for the features based on binary node importance (NI), as in (8) and (9). The feature importance of the RF was calculated in (10). In the final step, features that are of high importance are selected.

$$NI_j = w_j I_j - w_j^{left} I_j^{left} - w_j^{right} I_j^{right} \tag{8}$$

$$FI_i = \frac{\sum_{j=1}^{i} NI_j}{\sum_{k=1}^{n} NI_k} \tag{9}$$

$$RF(FI_i) = \frac{1}{T}\sum_{j=1}^{T}\left(\frac{FI_i}{\sum_{k=1}^{f} FI_k}\right)_j \tag{10}$$

where NI_j is the importance of node j-th, w_j is the weighted sampling on node j-th, I_j is the impurity on node j-th (left and right node), i is feature i-th, n is the total number of nodes, f is the total number of features, and T is the total number of trees.

3.3. *Class Reduction*

The class reduction (CR) presented in this work is a solution to the problem of an excessive number of classes, but a grouping of classes can be achieved. Initially, nine classes were defined: C1, C2, C3, F1, F2, F3, S1, S2, and S3 as shown in Table 1. However, there are some cases where the quality of the water is suitable for various farms at the same time. Alternatively, in some cases, the same water quality may be suitable for one farm but not for another. For example, the water quality suitable for crop farming is classified as class C1 (normal level); conversely, it is classified as class F2 (caution level) for fish farming. It may be possible for any water quality to be classified into three different classes based on farming activities under the same model.

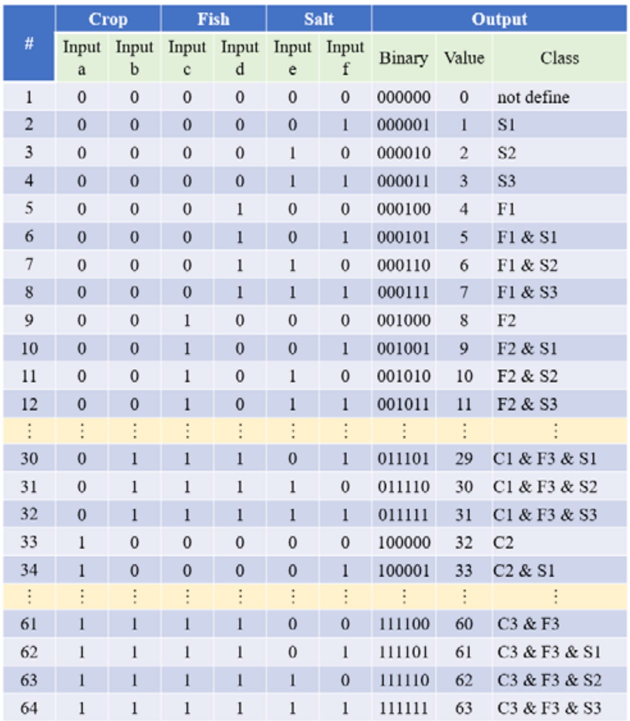
3.3.1. *Binary Concatenation*

Considering the maximum number of classes per farm, there are four levels of possibilities: normal, caution, critical, and unclassified (null). The probability can be represented as two bits in Table 2 for various farms.

Table 2. The probability of class labeled for various farms

Farm	Input 1	Input 2	Class
Crop	0	0	null
	0	1	C1
	1	0	C2
	1	1	C3
Fish	0	0	null
	0	1	F1
	1	0	F2
	1	1	F3
Salt	0	0	null
	0	1	S1
	1	0	S2
	1	1	S3

When considering the probability of class occurring within the water quality of the same sample, it is 2N, where N refers to the number of farm types. The total class output probability for the three farms was 64 when concatenating the binary input of the three farm types, as shown in Figure 3. However, the probabilities occurring in this research are only 63 values because they must have at least one result class and are not null for all farms (see 'not define' a class in Figure 3).



#	Crop		Fish		Salt		Output		
	Input a	Input b	Input c	Input d	Input e	Input f	Binary	Value	Class
1	0	0	0	0	0	0	000000	0	not define
2	0	0	0	0	0	1	000001	1	S1
3	0	0	0	0	1	0	000010	2	S2
4	0	0	0	0	1	1	000011	3	S3
5	0	0	0	1	0	0	000100	4	F1
6	0	0	0	1	0	1	000101	5	F1 & S1
7	0	0	0	1	1	0	000110	6	F1 & S2
8	0	0	0	1	1	1	000111	7	F1 & S3
9	0	0	1	0	0	0	001000	8	F2
10	0	0	1	0	0	1	001001	9	F2 & S1
11	0	0	1	0	1	0	001010	10	F2 & S2
12	0	0	1	0	1	1	001011	11	F2 & S3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	0	1	1	1	0	1	011101	29	C1 & F3 & S1
31	0	1	1	1	1	0	011110	30	C1 & F3 & S2
32	0	1	1	1	1	1	011111	31	C1 & F3 & S3
33	1	0	0	0	0	0	100000	32	C2
34	1	0	0	0	0	1	100001	33	C2 & S1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
61	1	1	1	1	0	0	111100	60	C3 & F3
62	1	1	1	1	0	1	111101	61	C3 & F3 & S1
63	1	1	1	1	1	0	111110	62	C3 & F3 & S2
64	1	1	1	1	1	1	111111	63	C3 & F3 & S3

Figure 3. An example of class output probability for three types of farms

3.3.2. Adding Feature by Class Grouping

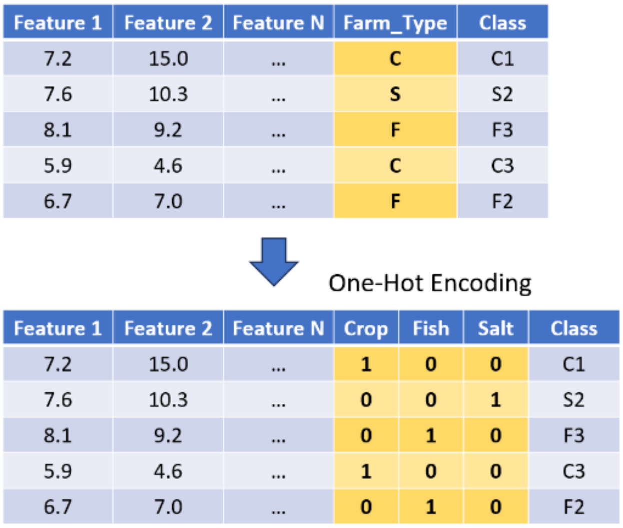
According to Figure 3, with an increase in the *N* number of farm types, class probabilities increase to exponential 2^N. These will significantly deteriorate the model's performance as many classes increase. The approach to reducing the number of classes in this research is to group the classes by farm types and break them into a feature. Therefore, a feature 'farm type' was added to the dataset. This farm type feature has values including 'C,' 'F,' and 'S' for the crop, fish, and salt farms, respectively. Then the number probability of class output was decreased to N^2, as same as the number of classes defined, as in Table 1. After adding a type of farm feature to the dataset, the data will be multiplied by N types of farms. The final dataset will contain 6,315 records (2,105 records × 3 farm types).

3.4. Model Development

The water quality classification model developed in this work consists of two parts: multi-layer perceptron neural network-based modeling and multi-class support vector machine classifying. However, there are three processes of the model development as follows.

3.4.1. One-Hot Encoding

The dataset that has undergone the feature selection process with RFFIS is then used to develop a water quality classification model. However, the attribute 'farm type' is the qualitative data that cannot be directly processed in the neural network. A one-hot encoding method will convert this feature to binarization. This feature will be converted into three attributes according to the category in the feature data ('C,' 'F,' and 'S'), as shown in Figure 4.



Feature 1	Feature 2	Feature N	Farm_Type	Class
7.2	15.0	...	C	C1
7.6	10.3	...	S	S2
8.1	9.2	...	F	F3
5.9	4.6	...	C	C3
6.7	7.0	...	F	F2

One-Hot Encoding

Feature 1	Feature 2	Feature N	Crop	Fish	Salt	Class
7.2	15.0	...	1	0	0	C1
7.6	10.3	...	0	0	1	S2
8.1	9.2	...	0	1	0	F3
5.9	4.6	...	1	0	0	C3
6.7	7.0	...	0	1	0	F2

Figure 4. The one-hot encoding for qualitative data

3.4.2. Multi-Layer Perceptron Neural Network-Based Modeling

The Multi-Layer Perceptron model generally consists of an input, hidden, and output layer. As for hidden layers, there can be more than one layer within a single MLPNN model. Adding more hidden layers increases the capacity of the network to learn intricate relationships and patterns in the data. Deeper networks can model highly nonlinear and hierarchical representations, enabling them to capture more complex functions. In addition, more neurons or units for each hidden layer allow for finer feature extraction. The MLPNN model in this research consists of eight layers, including an input layer, five hidden layers, one dropout layer, and one output layer.

Let *f* be the number of input features (eight features, after feature selection and one-hot encoding processes), and *c* be the total number of classes (nine classes). Each hidden layer has a number of neurons that were calculated in (11), (12), and (13).

$$N_1 = N_2 = 2\times(f+c) \tag{11}$$

$$N_3 = N_4 = f+c \tag{12}$$

$$N_5 = \left\lceil \frac{2}{3} f \right\rceil + c \tag{13}$$

where N_1, N_2, N_3, N_4, and N_5 are the number of neurons in the first to the fifth hidden layer, respectively.

In the Rectified Linear Unit (ReLU), the activation function was set to each neuron in the hidden layer. In addition, each neuron calculates the summation of the weights obtained from each input, as in (14) [36] and illustrated in Figure 5.

$$y = f(\sum_{i=1}^{n} w_i x_i + b) \tag{14}$$

where n is the number of inputs, w_i is the weight, x_i is the input, b is the bias, and f is the activation function.

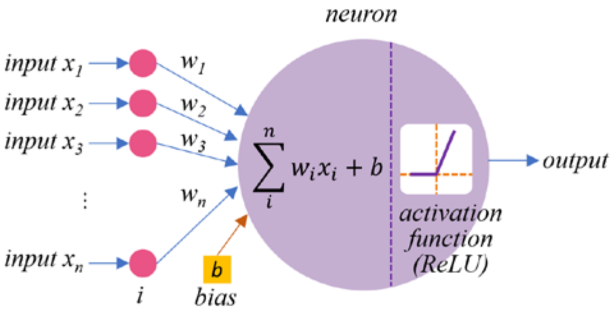


Figure 5. The proposed deep neural network framework

The MLPNN-based model sets the parameters for each layer, as shown in Table 3.

Table 3. The parameter for MLPNN-based model's layers

Layer	No. of input /neurons	Activation Function/ Classifier	No. of trainable parameters
Input	8	-	0
1st Hidden	34	ReLU	306
2nd Hidden	34	ReLU	1190
3rd Hidden	17	ReLU	595
4th Hidden	17	ReLU	306
5th Hidden	15	ReLU	270
Dropout (0.5)	-	-	0
Output	9	Softmax, OvO SVM, OvA SVM	144

Some hyperparameters were defined for the MLPNN-based model: batch size = 100, epoch = 500, and Adam learning rate = 0.001.

3.4.3. Multi-Class Support Vector Machine Classifying

Usually, most studies define a classifier with the Softmax activation function for the output layer in a neural network. However, the MCSVM was applied to the output layer of the MLPNN-based model in this work. There are two popular MCSVM approaches: One-versus-One (OvO) and One-versus-All (OvA), which are based on a binary SVM classifier that finds the decision hyperplane in (15) [37].

$$w^T x + b = 0 \tag{15}$$

where w^T is the decision vector in training dataset T, x is the input data point, and b is the displacement term.

Assume N is the total number of classes. The model will be trained 0.5×N×(N-1) binary classifiers for OvO SVM, while the OvA SVM involves training only N binary classifiers. This work applied OvO SVM and OvA SVM as classifiers and compared the model's efficiency between OvO SVM, OvA SVM, and Softmax activation function. Nevertheless, twelve models were developed with and without feature selection (FS) and class reduction (CR) to improve the water quality classification model. The dataset has fifteen features in this work. If it is processed using RFFIS, it will leave five features. In the case of processing with the class reduction approach, three additional features are obtained using one-hot encoding. All developed models were defined in Table 4.

Table 4. The developed MLPNN-based model

Model name	Classifier	Feature selection	Class reduction	No. of Input	No. of Classes
Softmax	Softmax	No	No	15	63
Softmax+cr		No	Yes	18	9
Softmax+fs		Yes	No	5	63
Softmax+fs+cr		Yes	Yes	8	9
OvO SVM	OvO SVM	No	No	15	63
OvO SVM+cr		No	Yes	18	9
OvO SVM+fs		Yes	No	5	63
OvO SVM+fs+cr		Yes	Yes	8	9
OvA SVM	OvA SVM	No	No	15	63
OvA SVM+cr		No	Yes	18	9
OvA SVM+fs		Yes	No	5	63
OvA SVM+fs+cr		Yes	Yes	8	9

3.5. Model Evaluation

Twelve MLPNN-based models were evaluated for performance in the training and validation process with accuracy (Acc), precision (Prec), sensitivity (Sens), F1-score, and MAE. These efficiencies were formulated from (16) to (19) [38], [39], [40].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{18}$$

$$F1\text{-}score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{19}$$

Where TP is the number of positive instances that were correctly classified as positive, TN is the number of negative instances that were correctly classified as negative, FP is the number of negative instances that were mistakenly classified as positive, and FN is the number of positive instances that were mistakenly classified as negative.

During the training process, the categorical cross-entropy loss was used to evaluate the model's efficiency for multi-class in neural networks, which is formulated in (20) [40].

$$Cross\text{-}entropy\ Loss = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} y_m^k \log(h_\theta(x_m, k)) \tag{20}$$

where M is the number of instances, K is the number of classes, y_m^k is the target label for instance m-th corresponding to class k-th, h_θ is the model with weights θ, and x_m is the input of instance m-th.

4. Results

The results of improving the water quality classification model for various farms are as follows:

4.1. The Result of Feature Selection

The fifteen features, including WT, pH, DO, EC, salinity, BOD, COD, turbidity, NH3, Pb, Cu, Cd, Hg, As, and flow rate, were processed by applying the RFFIS between Gini importance (or MDI) and permutation importance (or MDA) techniques. The results showed that both techniques yielded consistent feature selection results, with the top five feature importance being DO, salinity, pH, EC, and WT, respectively. These five features were selected as the MLPNN-based model's input features. In addition, the feature importance values for the top five features with the MDA method are higher than the MDI method. For example, the feature 'DO' has the highest mean importance values of 0.2140467 and 0.1767013 for MDA and MDI, respectively. On the contrary, the remaining features of MDA are gradually approaching zero importance which is closer to the zero than the MDI method. This makes the decision to remove features that are least relevant for the prediction easier. The mean and standard deviation values of importance for both methods using RFFIS are shown in Table 5. The feature importance values of MDI and MDA can be compared, as shown in Figure 6.

Table 5. The result of feature selection using RFFIS

Feature	**Mean Decrease Impurity**		**Mean Decrease Accuracy**	
	Mean	**SD**	**Mean**	**SD**
DO	0.1767013	0.0148484	0.2140467	0.0158851
Salinity	0.1538556	0.0165725	0.1922613	0.0148018
pH	0.1435543	0.0158851	0.1765074	0.0148484
EC	0.1299867	0.0136656	0.1645689	0.0165725
WT	0.1119544	0.0148018	0.1502005	0.0136656
NH3	0.0420619	0.0079484	0.0202241	0.0086382
Turbidity	0.0409287	0.0094288	0.0199232	0.0077280
BOD	0.0296924	0.0086454	0.0153705	0.0079484
COD	0.0296208	0.0088548	0.0153612	0.0081829
Pb	0.0259028	0.0086382	0.0103348	0.0091971
Cd	0.0258209	0.0091971	0.0088101	0.0079658
Cu	0.0240247	0.0081829	0.0064959	0.0088548
Hg	0.0236274	0.0079658	0.0025531	0.0094288
As	0.0216871	0.0076391	0.0020926	0.0086454
Flow rate	0.0205805	0.0077280	0.0012491	0.0076391

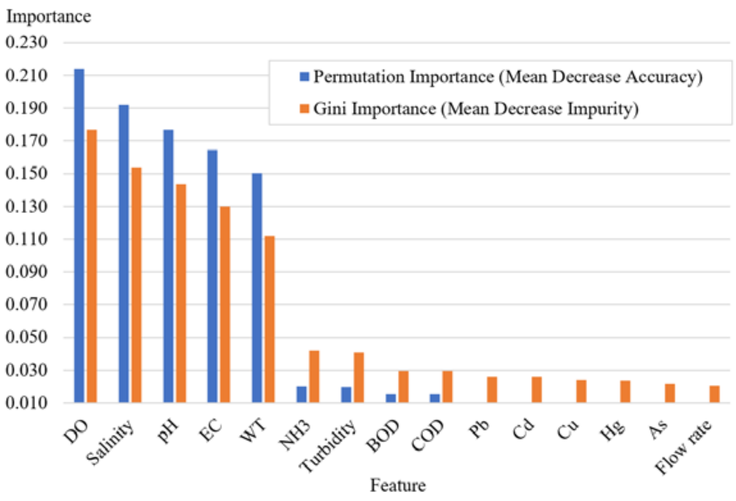


Figure 6. The comparison of feature importance between MDI and MDA

4.2. The Efficiency Result of the Model Evaluation

All developed models were evaluated for performance during the training and validation processes. The results of the model efficiency evaluation are shown in Table 6.

Table 6. The efficiency results of the models

Model	Acc	Prec	Sens	F1-score	MAE
OvA SVM+fs+cr	93.73	92.42	95.97	95.30	0.07033
OvO SVM+fs+cr	93.46	92.52	95.36	94.88	0.07345
Softmax+fs+cr	93.32	92.61	94.99	94.63	0.07512
OvA SVM+cr	92.51	92.69	93.38	93.47	0.08444
OvO SVM+cr	92.38	92.74	93.10	93.28	0.08591
Softmax+cr	92.21	92.79	92.73	93.02	0.08793
OvO SVM+fs	91.80	92.40	92.35	92.65	0.09257
OvA SVM+fs	91.67	92.34	92.18	92.51	0.09402
Softmax+fs	91.56	92.47	91.85	92.30	0.09533
OvO SVM	90.72	92.23	90.51	91.26	0.10494
OvA SVM	90.64	92.28	90.32	91.13	0.10586
Softmax	90.45	92.15	90.08	90.92	0.10802

According to Table 6, the developed MLPNN-based OvA SVM using feature selection and class reduction (OvA SVM+fs+cr) has the highest efficiency compared to other models in this work. The efficiency values of this model are accuracy of 93.73%, precision of 92.42%, sensitivity of 95.97%, F1-score of 95.30%, and MAE of 0.07033. The next five most efficient models were OvO SVM+fs+cr, Softmax+fs+cr, OvA SVM+cr, OvA SVM+cr, and Softmax+cr, with an accuracy of 93.46%. , 93.32%, 92.51%, 92.38%, and 92.21% respectively. Further, the MLPNN-based model applied the Softmax activation has an efficiency less than the OvA SVM and OvO SVM classifiers.

In addition, it was found that the MLPNN-based models that applied both feature selection and class reduction techniques had an efficiency higher than the model that applied the feature selection or did not apply both these techniques. For example, the OvA SVM+cr model has an accuracy of 92.51%, greater than the OvA SVM+fs model, whose accuracy was 91.67%.

Furthermore, all models were assessed for loss while training the models, with the results shown in Table 7. It was found that the order of model performance in Table 7 was consistent with the results in Table 6. The most efficient model is OvA SVM+fs+cr, with a validation accuracy of 89.14%, a training loss of 0.06942, and a validation loss of 0.25172. The validation accuracy of this model is less than the training accuracy, around 4.59%. The comparison of training and validation related to an iteration of training (500 epochs) for the MLPNN-based OvA SVM+fs+cr model is illustrated in Figure 7 and Figure 8.

Table 7. The training and validation loss of the models

Model	Accuracy		Loss	
	Training	Validation	Training	Validation
OvA SVM+fs+cr	93.73	89.14	0.06942	0.25172
OvO SVM+fs+cr	93.46	88.81	0.07240	0.28749
Softmax+fs+cr	93.32	88.25	0.07398	0.33773
OvA SVM+cr	92.51	87.72	0.08292	0.45331
OvO SVM+cr	92.38	87.30	0.08432	0.48554
Softmax+cr	92.21	87.16	0.08625	0.51064
OvO SVM+fs	91.80	86.75	0.09080	0.72164
OvA SVM+fs	91.67	87.11	0.09221	0.64639
Softmax+fs	91.56	86.43	0.09343	0.76136
OvO SVM	90.72	86.22	0.10272	0.97402
OvA SVM	90.64	86.13	0.10360	0.87897
Softmax	90.45	85.07	0.10570	1.32755

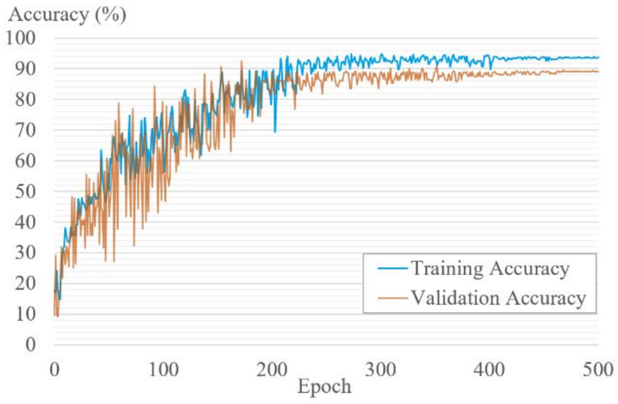


Figure 7. The training and validation accuracy of the OvA SVM+fs+cr model

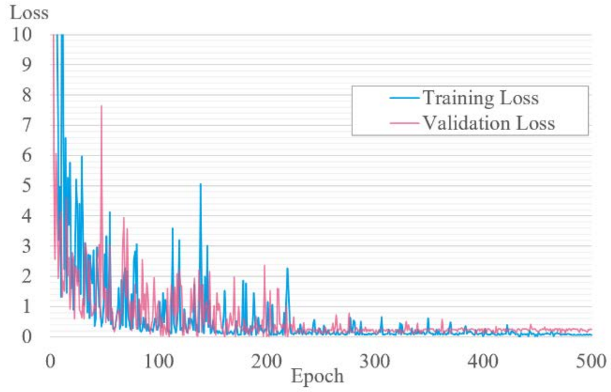


Figure 8. The training and validation loss of the OvA SVM+fs+cr model

5. Conclusion and Discussion

Water is essential to the livelihoods of farmers and villagers, including use in various farming processes from the past to the present. In particular, areas near or adjacent to the coast will always be affected by seawater. It has resulted in seawater salinity infiltrating various farming areas, including crop farms, fish farms, and salt farms. Further, the water quality parameters such as dissolved oxygen, pH, electrical conductivity, and water temperature affect these farms. Especially, areas with overlapping water ecosystems make it difficult to distinguish the appropriate water quality for various farms and use water resources within the same source. Therefore, this research aims to develop and improve the efficiency of water quality classification models for various farms using feature-based MLPNN and multi-class SVM. The MLPNN-based model architecture includes an input layer, five hidden layers, one dropout layer, and an output layer. The results showed that the processed dataset using RFFIS gave better model performance than the dataset that did not use this technique. With the Gini importance and permutation importance methods, it is found that DO is the feature that has the most significant relationship to the result class.

It aligns with Suwadi's work [42]. The following features are salinity, pH, EC, and WT. Further, the proposed class reduction for classes that can be grouped can significantly improve the model's efficiency by adding attributes from those grouped classes and reducing the number of co-occurring classes from an instance. Considering the distinction between the Softmax activation function, OvO SVM, and OvA SVM, it was found that using OvO SVM and OvA SVM, which are multi-class SVM, provides a higher model's efficiencies than the Softmax activation function defined within the output layer of MLPNN-based model which is the baseline for comparison. Moreover, MLPNN-based models using the OvO SVM classifier have higher model performance than the OvA SVM classifier in cases where the model does not use class reduction techniques. This is because the OvO SVM classifier performs better when using smaller datasets, such as the iris and wine datasets, according to López and Maldonado [43]. However, with a larger dataset, OvO SVM tends to become less efficient due to the increased training time of multiple classifiers. Overall, the findings indicate that a model developed with a Multi-Layer Perceptron Neural Network combined with a One-versus-All Support Vector Machine using datasets processed by Random Forest Feature Importance Selection and class reduction techniques can classify water quality with an accuracy of 93.73%.

We studied previous research studies that have focused on models developed from one classifier for machine learning modeling and then compared the efficiency of those models. For the best way to optimize the water quality prediction model, this work combines the neural network with the SVM to aid in water quality classification in the final step of the model. The improved model can be used for classifying water quality for crop, fish, and salt farms with different water quality requirements in an area that overlaps the water ecosystem and is affected by seawater. Although the model uses an artificial neural network with five hidden layers which adds the ability to capture patterns during the feature extraction step, due to the increased number of parameters and computations involved, deeper networks generally require more computational resources to train and evaluate. Hardware and parallel computing advancements have made training deep MLPNN more feasible. Several parameters could be improved for future work, such as hidden layer architectures, learning rate, batch size, and number of dropout layers and their rate. This experiment applied a dropout rate of 0.5 which is suitable and improves the model's accuracy.

For future work, the hyperparameters could fine-tune and optimize the suitable values for the neural network model.

Besides, the feature extraction output from the last hidden layer could be transferred learning to other models for improving the efficiency of water quality classification. The next model could use the Long Short-Term Memory (LSTM) to predict the real-time water quality and upcoming. In addition, the collection area of water quality sampling should be expanded, and the model must cover a wide range of farm types. The newly developed model will be used to create a mobile application for villagers, farmers, and related officials to be notified of the water quality in the current situation in advance to take corrective actions and manage the water quality appropriately.

Acknowledgments

The authors are grateful to the Institute for Research and Development, the Faculty of Science and Technology, Suan Sunandha Rajabhat University, and the Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, who supported and gave this research opportunity. The research conducted in this study received support from the national budget for the year 2023, provided by Suan Sunandha Rajabhat University. The project was carried out under the permission/contract number 11803/2023, with a research ethics certificate number COE.1-039/2023.

References:

[1]. Kawahata, H., Fujita, K., Iguchi, A., Inoue, M., Iwasaki, S., Kuroyanagi, A., Maeda, A., Manaka, T., Moriya, K., Takagi, H., Toyofuku, T., Yoshimura, T., & Suzuki, A. (2019). Perspective on the response of marine calcifiers to global warming and ocean acidification—Behavior of corals and foraminifera in a high CO2 world "hot house". *Progress in Earth and Planetary Science*, *6*(1). Doi:10.1186/s40645-018-0239-9

[2]. Kim, J.-H., Kim, N., Moon, H., Lee, S., Jeong, S. Y., Diaz-Pulido, G., Edwards, M. S., Kang, J.-H., Kang, E. J., Oh, H.-J., Hwang, J.-D., & Kim, I.-N. (2020). Global warming offsets the ecophysiological stress of ocean acidification on temperate crustose coralline algae. *Marine Pollution Bulletin*, *157*. Doi:10.1016/j.marpolbul.2020.111324

[3]. Paltán, H. A., Pant, R., Plummer Braeckman, J., & Dadson, S. J. (2021). Increased water risks to global hydropower in 1.5 °C and 2.0 °C Warmer Worlds. *Journal of Hydrology*, 599. Doi:10.1016/j.jhydrol.2021.126503

[4]. Li, H., Cui, Z., Cui, H., Bai, Y., Yin, Z., & Qu, K. (2023). Hazardous substances and their removal in recirculating aquaculture systems: A review. *Aquaculture*, *569*. Doi:10.1016/j.aquaculture.2023.739399

[5]. Zafar, M. A., Talha, M. A., & Rana, M. M. (2021). Effect of biofloc technology on growth performance, digestive enzyme activity, proximate composition, and hematological parameters of Asian stinging catfish (Heteropneustes fossilis). *Journal of Applied Aquaculture*, *34*(3), 755–773. Doi:10.1080/10454438.2021.1957053

[6]. Ayesha Jasmin, S., Ramesh, P., & Tanveer, M. (2022). An intelligent framework for prediction and forecasting of dissolved oxygen level and biofloc amount in a shrimp culture system using machine learning techniques. *Expert Systems with Applications*, 199. Doi:10.1016/j.eswa.2022.117160

[7]. Davis, T. R., Larkin, M. F., Forbes, A., Veenhof, R. J., Scott, A., & Coleman, M. A. (2022). Extreme flooding and reduced salinity causes mass mortality of nearshore kelp forests. *Estuarine, Coastal and Shelf Science*, *275*. Doi:10.1016/j.ecss.2022.107960

[8]. Stoyanova, E. (2023). Remote sensing for flood inundation mapping using various processing methods with Sentinel-1 and Sentinel-2. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 339–346. Doi:10.5194/isprs-archives-XLVIII-M-1-2023-339-2023

[9]. Walukow, A., & Sukarta, I. N. (2021). Analysis of carrying capacity and water pollution in the Simporo Strait area after the flash flood. *Ecological Engineering & Environmental Technology*, *22*(3), 120–128. Doi:10.12912/27197050/135528

[10]. Prakash, S. (2021). Impact of climate change on aquatic ecosystem and its biodiversity: An overview. *International Journal Biological Innovations*, *3*(2). Doi:10.46505/ijbi.2021.3210

[11]. Goodrich, H. R., Berry, A. A., Montgomery, D. W., Davison, W. G., & Wilson, R. W. (2022). Fish feeds supplemented with calcium-based buffering minerals decrease stomach acidity, increase the blood alkaline tide and cost more to digest. *Scientific Reports*, *12*(1). Doi:10.1038/s41598-022-22496-3

[12]. Harpaz, S., Hakim, Y., Slosman, T., & Eroldogan, O. T. (2005). Effects of adding salt to the diet of Asian sea bass Lates calcarifer reared in fresh or salt water recirculating tanks, on growth and brush border enzyme activity. *Aquaculture*, *248*, 315–324. Doi:10.1016/j.aquaculture.2005.03.007

[13]. de Luna Souto, A. G., Cavalcante, L. F., de Melo, E. N., Cavalcante, Í. H. L., da Silva, R. Í. L., de Lima, G. S., Gheyi, H. R., Pereira, W. E., de Paiva Neto, V. B., de Oliveira, C. J. A., & de Oliveira Mesquita, F. (2023). Salinity and mulching effects on nutrition and production of grafted sour passion fruit. *Plants*, *12*(5). Doi:10.3390/plants12051035

[14]. Surendran, S. N., Jayadas, T. T. P., Tharsan, A., Thiruchenthooran, V., Santhirasegaram, S., Sivabalakrishnan, K., Raveendran, S., & Ramasamy, R. (2020). Anopheline bionomics, insecticide resistance and transnational dispersion in the context of controlling a possible recurrence of malaria transmission in Jaffna city in northern Sri Lanka. *Parasites & Vectors*, *13*(1). Doi:10.1186/s13071-020-04037-x

[15]. Kim, B., Lee, S.-m., Kang, S.-h., Jeong, M.-s., Gim, G. H., Park, J., & Lim, C. (2020). Aquavoltaic system for harvesting salt and electricity at the salt farm floor: Concept and field test. *Solar Energy Materials and Solar Cells, 204*. Doi:10.1016/j.solmat.2019.110234

[16]. Hassan, H. U. et al. (2024). Growth performance and survivability of the Asian seabass Lates calcarifer (Bloch, 1790) reared under hyper-saline, hypo-saline and freshwater environments in a closed aquaculture system. Brazilian *Journal of Biology*, 84. Doi:10.1590/1519-6984.254161

[17]. Promput, S., Maithomklang, S., & Panya-isara, C. (2023). Design and analysis performance of IoT-based water quality monitoring system using LoRa technology. *TEM Journal*, *12*(1), 29–35. Doi:10.18421/tem121-04

[18]. Jha, B. K. (2020). Cloud-based smart water quality monitoring system using IoT sensors and machine learning. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3403–3409. Doi:10.30534/ijatcse/2020/141932020

[19]. Chowdury, M. S. U., Emran, T. B., Ghosh, S., Pathak, A., Alam, M. M., Absar, N., Andersson, K., & Hossain, M. S. (2019). IoT based real-time river water quality monitoring system. *Procedia Computer Science*, *155*, 161–168. Doi:10.1016/j.procs.2019.08.025

[20]. Salahat, E., & Qasaimeh, M. (2017). Recent advances in features extraction and description algorithms: a comprehensive survey. In *Proceeding of 2017 IEEE International Conference on Industrial Technology (ICIT)*, Toronto, ON, Canada, 22-25 March 2017, 1059–1063. Doi:10.1109/icit.2017.7915508

[21]. Uyun, S., & Sulistyowati, E. (2020). Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(4). Doi:10.11591/ijece.v10i4.pp4331-4339

[22]. Bakar, A. A., Hamdan, R., & Sani, N. S. (2020). Ensemble learning for multidimensional poverty classification. *Sains Malaysiana*, *49*(2), 447–459. Doi:10.17576/jsm-2020-4902-24

[23]. Alhutaish, R., & Omar, N. (2017). Feature Selection for Multi-label Document Based on Wrapper Approach through Class Association Rules. *International Journal on Advanced Science, Engineering and Information Technology*, *7*(2). Doi:10.18517/ijaseit.7.2.1040

[24]. Kushwaha, N. L., Rajput, J., Suna, T., Sena, D. R., Singh, D. K., Mishra, A. K., Sharma, P. K., & Mani, I. (2023). Metaheuristic approaches for prediction of water quality indices with relief algorithm-based feature selection. *Ecological Informatics*, *75*. Doi:10.1016/j.ecoinf.2023.102122

[25]. Malik, H., & Yadav, A. K. (2021). A novel hybrid approach based on relief algorithm and fuzzy reinforcement learning approach for predicting wind speed. *Sustainable Energy Technologies and Assessments*, *43*. Doi:10.1016/j.seta.2020.100920

[26]. Lap, B. Q., Phan, T.-T.-H., Nguyen, H. D., Quang, L. X., Hang, P. T., Phi, N. Q., Hoang, V. T., Linh, P. G., & Hang, B. T. T. (2023). Predicting water quality index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. *Ecological Informatics*, *74*. Doi:10.1016/j.ecoinf.2023.101991

[27]. Ilić, M., Srdjević, Z., & Srdjević, B. (2022). Water quality prediction based on Naïve Bayes algorithm. *Water Science and Technology*, *85*(4), 1027–1039. Doi:10.2166/wst.2022.006

[28]. Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water quality prediction using KNN imputer and Multilayer Perceptron. *Water*, *14*(17). Doi:10.3390/w14172592

[29]. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, *11*(11). Doi:10.3390/w11112210

[30]. Leong, W. C., Bahadori, A., Zhang, J., & Ahmad, Z. (2019). Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM). *International Journal of River Basin Management*, *19*(2), 149–156. Doi:10.1080/15715124.2019.1628030

[31]. Nair, J. P., & Vijaya, M. S. (2022). River water quality prediction and index classification using machine learning. *Journal of Physics: Conference Series*, *2325*(1). Doi:10.1088/1742-6596/2325/1/012011

[32]. Aish, A. M., Zaqoot, H. A., Sethar, W. A., & Aish, D. A. (2023). Prediction of groundwater quality index in the Gaza coastal aquifer using supervised machine learning techniques. *Water Practice and Technology*, *18*(3), 501-521. Doi:10.2166/wpt.2023.028

[33]. Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*(2). Doi:10.1007/s42452-021-04148-9

[34]. Bouke, M. A., Abdullah, A., Alshatebi, S. H., Abdullah, M. T., & Atigh, H. E. (2023). An intelligent DDoS attack detection tree-based model using Gini index feature selection method. *Microprocessors and Microsystems*, *98*. Doi:10.1016/j.micpro.2023.104823

[35]. Algehyne, E. A., Jibril, M. L., Algehainy, N. A., Alamri, O. A., & Alzahrani, A. K. (2022). Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia. *Big Data and Cognitive Computing*, *6*(1). Doi:10.3390/bdcc6010013

[36]. Kasasbeh, B., Aldabaybah, B., & Ahmad, H. (2022). Multilayer perceptron artificial neural networks-based model for credit card fraud detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1). Doi:10.11591/ijeecs.v26.i1.pp362-373

[37]. Meng, L., & Wu, C.-H. (2022). The promotion effect of the improved ISCA model on the application of accounting informatization in small- and medium-sized enterprises in the cloud computing environment. *Mobile Information Systems*, *2022*, 1–13. Doi:10.1155/2022/4228178

[38]. Nuanmeesri, S. (2022). Development of community tourism enhancement in emerging cities using gamification and adaptive tourism recommendation. *Journal of King Saud University - Computer and Information Sciences*, *34*(10), 8549–8563. Doi:10.1016/j.jksuci.2021.04.007

[39]. Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S., & Kadmateekarun, P. (2022). Improving Dropout Forecasting during the COVID-19 Pandemic through Feature Selection and Multilayer Perceptron Neural Network. *International Journal of Information and Education Technology*, *12*(9), 851–857. Doi:10.18178/ijiet.2022.12.9.1693

[40]. Kaddoura, S. (2022). Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability*, *14*(18). Doi:10.3390/su141811478

[41]. Ho, Y., & Wookey, S. (2020). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. *IEEE Access*, *8*, 4806–4813. Doi:10.1109/access.2019.2962617

[42]. Suwadi, N. A., Derbali, M., Sani, N. S., Lam, M. C., Arshad, H., Khan, I., Kim, K.-I., & Shuja, J. (2022). An optimized approach for predicting water quality features based on machine learning. *Wireless Communications and Mobile Computing*, *2022*, 1–20. Doi:10.1155/2022/3397972

[43]. López, J., & Maldonado, S. (2016). Multi-class second-order cone programming support vector machines. *Information Sciences*, *330*, 328–341. Doi:10.1016/j.ins.2015.10.016