

## Research Article

# An Optimized Approach for Predicting Water Quality Features Based on Machine Learning

Nur Afyfh Suwadi,<sup>1</sup> Morched Derbali,<sup>2</sup> Nor Samsiah Sani ,<sup>3</sup> Meng Chun Lam,<sup>1</sup> Haslina Arshad,<sup>4</sup> Imran Khan ,<sup>5</sup> and Ki-Il Kim <sup>6</sup>

<sup>1</sup>A Mixed Reality and Pervasive Computing Lab, Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

<sup>2</sup>Faculty of Computing and Information Technology (FCIT), King Abdulaziz University (KAU), Jeddah, Saudi Arabia

<sup>3</sup>Center for Artificial Intelligence Technology, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

<sup>4</sup>Institute of IR4.0 (IIR4.0), Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

<sup>5</sup>Department of Electrical Engineering, University of Engineering and Technology Peshawar, Pakistan

<sup>6</sup>Department of Computer Science and Engineering, Chungnam National University, Daejeon 34134, Republic of Korea

Correspondence should be addressed to Nor Samsiah Sani; [norsamsiahsani@ukm.edu.my](mailto:norsamsiahsani@ukm.edu.my), Imran Khan; [imran\\_khan@uetpeshawar.edu.pk](mailto:imran_khan@uetpeshawar.edu.pk), and Ki-Il Kim; [kikim@cnu.ac.kr](mailto:kikim@cnu.ac.kr)

Received 23 March 2022; Revised 7 August 2022; Accepted 17 August 2022; Published 9 September 2022

Academic Editor: Junaid Shuja

Copyright © 2022 Nur Afyfh Suwadi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traditionally, water quality is assessed using costly laboratory and statistical methods, rendering real-time monitoring useless. Poor water quality requires a more practical and cost-effective solution. The machine learning classification approach appears promising for rapid detection and prediction of water quality. Machine learning has been used successfully to predict water quality. However, research on machine learning for water quality index (WQI) prediction is generally lacking. Therefore, this research aims to identify the important features for the WQI, which necessitated the classification of numerous indicators. This study develops four machine learning models (Artificial Neural Network, Support Vector Machine, Random Forest, and Naïve Bayes) based on the WQI and chemical parameters. The Langat Basin in Selangor dataset from the Department of Environment of Malaysia trains and validates each machine learning model. Several data preprocessing tasks such as data cleaning and feature selection have been conducted on the raw dataset to ensure the quality of the training data. The performance of these machine learning algorithms is further rectified based on the selected features set by several feature selection strategies such as information gain, correlation, and symmetrical uncertainty. Each classifier is then optimized using different tuning parameters to achieve optimum values before comparing the output of the three classifiers against each other. The observational results have shown that the optimized Random Forest classifier with the WQI parameter selected by the information gain feature selection method achieved the highest performance. The experimental results show that the WQI parameters are more relevant in predicting the WQI than the other variables. Consequently, this result shows that parameter oxygen (DO) and biochemical oxygen demand (BOD) are important features for predicting WQI. The proposed model achieved reasonable accuracy with minimal parameters, indicating that it could be used in real-time water quality detection systems.

## 1. Introduction

The provision of sufficient water and the preservation of water resource quality, especially in river water systems, are crucial for the long-term development of countries worldwide [1, 2].

Water availability and quality, whether surface or groundwater, have declined due to several important factors, including population growth and urbanization [3]. Without a proper plan, urban development causes frequent environmental problems or increases human waste and activity, polluting

the environment [4]. As a result, we must devise the most efficient methods of maintaining water quality. Several parameters can be calculated to assess the consistency of water. The quality status could be graded based on the importance of a few selected attributes. The appropriate feature usage is one of the crucial components in deciding a classification method's performance grade [5]. Two processes related to features are extraction [5] and selection [6].

Water quality index (WQI) is Malaysia's most recent method of monitoring water quality developed by JAS (IKA-JAS). Since 1978, the DOE has been monitoring water quality, mainly to establish standards for detecting changes in water quality and identifying pollution sources. The IKA-JAS is also used to measure contamination levels and identify water quality according to the National Water Quality Standard's recommendations for water use (SKAN) [7]. Several national and international organizations, however, have established their WQI, such as the Weight Arithmetic WQI (WAWQI), National Sanitation Foundation WQI (NSFWQI), Canadian Council of Ministers of the Environment WQI (CCMEWQI), and Oregon WQI (OWQI), to name a few. These WQIs have been used to assess the water quality in a specific region [8]. Furthermore, these indices are frequently based on various number and water quality parameters instead of regional norms. WQIs have been certified to show annual cycles, spatial, and temporal variability in water quality, as well as the patterns in an effective and timely manner, even at low concentrations.

Although numerous feature selection methods have been proposed in recent years, the problem remains unresolved due to difficulties in selecting the minimum feature in the water quality of Lembangan Sungai Langat, which necessitates the classification of various indicators. If the model's accuracy percentage remains constant, the processing time to predict the class would be shorter with fewer features. As a result, it is preferable to classify using important variables rather than selecting all [9]. The study is experiencing general performance issues because of different indicators used to classify water quality. In machine learning, feature selection is a critical activity that aims to eliminate irrelevant and redundant features that can hinder performance [10]. The monitoring and estimation of water quality have sparked a lot of interest. Hence, different machine learning processes based on feature selection have been used to make such predictions. For instance, Yu and Bai [11] compared the intelligent optimization models applying the genetic algorithm [12], optimized the hyperparameter implementing particle swarm optimization as feature selection, and [13] used SARIMA-LSTM, a hybrid algorithm that integrates deep learning and conventional time series models as a multifeature prediction method to apply correlation as feature selection.

On the other hand, previous studies did not thoroughly investigate feature selection on DOE-WQI. There is a study of WQI using machine learning models on Malaysia region before; however, they are more focusing on WQI features only or certain features in WQI [14–16]. Meanwhile, these studies are conducting an experiment using all the 28 water quality features. The feature selection method will also be used for each feature. As a result, the primary purpose of this paper

was to examine the effect of advanced feature selection and classification algorithms on data prediction. Feature selection was thought to influence machine learning precision. The best feature selection approach will be used in this analysis, focusing on selecting subsets of features and robust classification algorithms to predict water quality data. Information gain, correlation, and symmetrical uncertainty, as well as an applied classification algorithm, were the feature selection techniques proposed for implementation. The performance of these techniques was compared to find the best functionality.

This study aimed to improve prediction efficiency by selecting the best feature subset from the fewest number of available features. As a result, the reduced number of features reduces the time required to predict the class. SMOTE sampling methods were applied to the raw dataset to ensure the accuracy of the training data. The SMOTE technique effectively drives the minority class's decision to become more general. Oversampling the minority class is one way to deal with unbalanced datasets. Duplicating instances in the minority class is the easiest way, but these examples do not provide any new information to the model. Instead, new examples can be created by synthesizing old ones [17]. Each classifier was optimized using different tuning parameters with 10-fold cross-validation and split test 70-30 to achieve the optimal values until the output of the three classifiers was compared. Four machine learning algorithms followed the feature selection stages: ANN, SVM, Random Forest, and Naïve Bayes. From the previous water quality studies, that model has been applied to a certain type of machine learning model, and they are being compared with another type of classifier [18–20]; still, the comparison between the ANN, SVM, Random Forest, and Naïve Bayes need to be explored further. Numerous prediction challenges were successfully addressed using ANN [21]. It is a model that is applied to time-series prediction in a variety of engineering applications as a very effective machine learning technique [22]. Meanwhile, SVMs are mainly used for classification. SVMs use data points plotted on a plane to visualize the data, defining a hyperplane between the classes and extending the margin to maximize the distinction between two classes, leading to fewer close inaccurate measurements [23]. Then, a random forest utilizes various base models on different subsets of the provided data and bases its judgments on all of the models. Decision trees serve as the basis of the random forest model, which combines its advantages with the efficiency of combining many models [24]. The Naïve Bayes classification algorithm is based on the independence of characteristic conditions and the Bayes theorem. When the target value is provided, attributes are taken to be conditionally independent of one another [22].

Finally, several statistical tests will be run to validate the effects of the chosen classifier. A paired corrected *t*-test was used to compare output between learning methods for both the cross-validated and percentage-split datasets. This test compared paired samples to determine if two learning models are statistically significantly different or if one is preferable to the other [25]. 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

to compare feature selection methods to the performance of the chosen classification algorithms. Finally, Section 4 illustrates the findings and conclusions, respectively.

With the increased use of machine learning techniques in hydrological applications, assessing the robustness, performance, and reliability of these ML models' predictions is crucial [26]. Machine learning is an area of artificial intelligence (AI) and computer science that focuses on using data and algorithms to understand how humans learn to continually improve accuracy [27]. Massive new computer and internet applications have exponentially produced vast quantities of data, including video, picture, text, speech, and data derived from social relationships and the emergence of the Internet of Things (IoT) and cloud computing. These data frequently have a high number of dimensions, making data interpretation and decision-making difficult. Feature selection processes high-dimensional data efficiently and improves learning performance in theory and practice [28, 29]. Feature selection refers to obtaining a subset from an original feature set based on a feature selection criterion that selects the appropriate parts of the dataset. It aids in data processing scale compression by removing unnecessary and irrelevant functions. Feature selection techniques can preprocess learning algorithms, and successful feature selection results can increase learning precision, shorten learning time, and simplify learning outcomes [28, 30].

Some experiments in feature selection of water quality were conducted in previous research. However, the studies focused on their primary goal and employed a variety of techniques. A study done by Uyun and Sulistyowati [6] in feature selection for multiple water quality used the SMOTE technique and bootstrapping to address the imbalanced class problem. The amount of imbalanced data in each category has been shown to affect the pattern recognition system's learning process. Rodriguez-Galiano et al. [31] also researched feature selection approaches for groundwater nitrate emission predictive modeling. They were primarily interested in determining the utility of various feature selection methods, identifying the primary sources of nitrate pollution, comprehending device dynamics, and mapping the probabilities of nitrate occurrence in groundwater above a threshold value. Their research, however, is limited to nitrate. Furthermore, Golabi et al. [1] investigated a feature selection approach in predicting biochemical oxygen demand (BOD). They compared classification models to determine which provided the highest prediction accuracy.

In contrast, we used ANN, SVM, Random Forest, and Naïve Bayes as classification techniques in this study. We compared the applications of these techniques to determine which had the highest accuracy. The selection was chosen from a previous study that applied a feature selection method and a prediction method for water quality. Rajaei et al. [32] reviewed studies that used artificial intelligence (AI)-based techniques to model river water quality. Among the various modeling approaches, ANN has proven to be a good and reliable technique for this reason. ANN is preferable to other AI models for simulating different types of water quality variables. The results showed that most of the water quality variables predicted using ANN had sufficient accuracies. As a result, we have included it in this paper to compare our feature selection process. According to Haghiabi et al. [33], when the accuracy of

the implemented models was evaluated using error indexes, SVM was found to be the most effective model in predicting water quality in Iran's critical watersheds. Meanwhile, Chama-kura and Saha [34] used SVM to employ an instance voting approach in their feature selection analysis. They devised a feature selection strategy that takes advantage of the limited instances' local data.

Besides that, Devi [35] used Random Forest as a classification technique to predict water quality in her study. She used Random Forest to predict the water quality in the study area's regions and classified them into three drinking-water categories: excellent, good, and poor. Meanwhile, Jaihind et al. [36] used Random Forest classification in an IoT-based real-time water quality monitoring system. His research compared Random Forest to J48 and Multilayer Perceptron and discovered that the former outperformed J48 and Multilayer Perceptron in detecting and classifying instances. Moreover, Li et al. [37] applied Random Forest to predict water BOD [38]. Their experiment showed that Random Forest models outperformed M5 models in BOD modeling. The findings indicated that Random Forest could aid decision-makers in managing emissions and meet quality standards in current and future conditions.

Random Forest was typically used to evaluate feature selection methods for object-based land cover mapping of remotely aerial vehicle imagery [39]. Most research focused on the uncertainty of feature selection for object-based classification rather than the per-pixel approach. Meanwhile, Bindra et al. [40] used Naïve Bayes to predict water classes in the Indian River. They stated that of all the classification techniques (Naïve Bayes, J48, SMO, and REPTree) applied on their dataset, Naïve Bayes produced the best results with the slightest error. Naïve Bayes was also used as a classifier [41]. The authors proposed a new population-based meta-heuristic approach for the gene selection problem that combined the SU and the HSA. The symmetrical uncertainty filter and harmony search algorithm wrapper were also proposed as a two-stage selection algorithm for gene selection problems in microarray datasets (SU-HSA). SU is used to help eliminate the genes that are not needed; meanwhile, HSA was chosen for the gene selection problem because it is a stochastic random search approach that is less parameter sensitive and solves the disadvantage of the building block theory of GAs by considering all existing solutions rather than only two (parents) in its reproduction [42].

Despite numerous experiments using those classifiers, there are still no comparison studies between RF, SVM, ANN, and Naïve Bayes for a feature selection in the water quality dataset. Therefore, in this analysis, we experimented with and tested those classifiers on a dataset. Previously, extensive research on the use of ANNs, SVM, Random Forest, and Naïve Bayes was conducted worldwide. However, research into determining algorithm accuracy is scarce. More specifically, a proper technique should be developed for sorting the experiment's input data. Feature selection algorithms were used to address these flaws to sort the input data to find the best option with the highest accuracy for selecting the best water quality feature. Indeed, this study aimed to (1) apply feature selection algorithms to select the appropriate inputs for the experiment, (2) employ SMOTE, (3) detect the

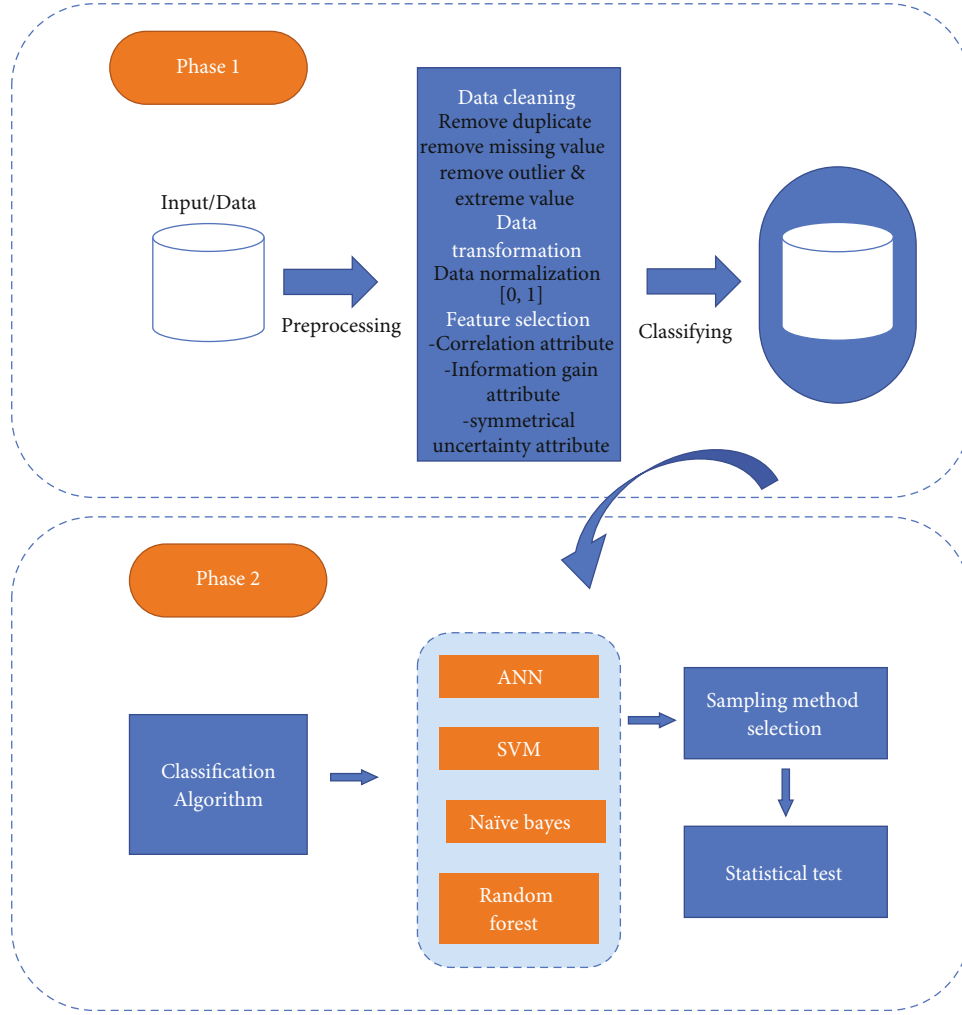


FIGURE 1: Experiment process flowchart.

classification algorithm's ability, and (4) use the paired corrected  $t$ -test to analyze their percentage of the corrected classifier with 0.05 two-tailed confidence.

## 2. Materials and Methods

An experimental design methodology was used in this study, following the experiment process processes shown in Figure 1.

**2.1. Dataset Description.** For this study, a water dataset from the Langat Basin in Selangor was used to identify the feature selection in water quality. The data was provided by the Department of Environment, Malaysia, the entity responsible for handling and monitoring the water quality in Malaysia. Table 1 summarizes the 29 attributes with 907 samples that were used in this study. The dataset used was collected between January 2012 and December 2018.

**2.2. Data Preprocessing.** Data preprocessing is transforming a dataset so that the information content is best exposed to the mining tool. Real-world data is inherently imperfect, unreliable, and prone to noise such as errors and outliers. As a result,

data preprocessing is required to ensure that the data is formatted for a specific miner tool and sufficient in a particular method [43, 44]. This analysis will perform several data preprocessing tasks, including cleaning, normalization, feature selection, and sampling methods. Data preprocessing also can be accomplished using a variety of data mining techniques.

The presence of noise in the attribute data would almost certainly affect its classification accuracy. If the data used has many attributes/parameters or features, it will undoubtedly affect the computation process [39]. Hence, the feature selection process becomes essential [6]. The preprocessing task was carried out using the Waikato Environment for Knowledge Analysis (WEKA) version 3.8 program in this report. WEKA, a java-based machine learning platform, was developed by the University of Waikato in New Zealand. WEKA is an open-source machine learning library that includes a variety of machine learning algorithms. It also has a range of visualization tools for data processing and forecasting [45].

**2.2.1. Data Cleaning.** The dataset contains some missing (null) values based on manual checking using filtering

TABLE 1: Detailed description.

	Attribute	Unit	Type	Description
National Water Quality Standard (13)	DO	mg/l	Real	Dissolved oxygen
	BOD	mg/l	Integer	Biochemical oxygen demand
	COD	mg/l	Integer	Chemical oxygen demand
	SS	mg/l	Integer	Total suspended solid
	pH	Unit	Real	pH
	NH3-NL	mg/l	Real	Ammoniacal nitrogen
	COND	$\mu S/cm$	Real	Electrical conductivity
	SAL	Ppt	Real	Salinity
	DS	mg/l	Real	Total dissolved solids
	TEMP	$^{\circ}C$	Real	Temperature
	TUR	NTU	Real	Turbidity
	E. coli	cfu/100ml	Integer	Fecal coliform
	Coliform	cfu/100 ml	Integer	Total coliform
All chemical content (27)	NO3	mg/l	Real	Nitrate
	Cl	mg/l	Real	Chlorine
	PO4	mg/l	Real	Phosphate
	As	mg/l	Real	Arsenic
	Hg	mg/l	Real	Mercury
	Cd	mg/l	Real	Cadmium
	Cr	mg/l	Real	Chromium
	Pb	mg/l	Real	Lead
	Zn	mg/l	Real	Zinc
	Ca	mg/l	Real	Calcium
	Fe	mg/l	Real	Iron
	K	mg/l	Real	Potassium
	Mg	mg/l	Real	Magnesium
	Na	mg/l	Real	Sodium
	WQI	—	Integer	Water quality index
	Class	—	Nominal	Class

functions in WEKA. As shown in Table 2, attributes with null values are manually replaced with the mean value [46, 47].

Besides, we replaced it with a whole number for the original value lower than scale, as described in Table 3.

The WEKA Explorer will automatically calculate descriptive statistics for numerical attributes, allowing us to identify data properties and noise, outliers, and extreme values. Next, Table 4 presents a brief overview of the data collection.

**2.2.2. Noisy Data.** Outliers may be presented in the dataset. An outlier is a dataset that deviates from most findings, most likely due to a different mechanism [44]. In this study, outliers were identified using WEKA's Interquartile Range. The Unsupervised Attribute Interquartile Range was used in WEKA to detect outliers before removing both the outliers and extreme values (Table 5).

**2.2.3. Normalization.** Normalization is a numeric-based scaling process in which the maximum and minimum values are commonly far apart, such as 1 and 10000. Due to normalization, the

TABLE 2: Replaced missing value with mean.

Attribute	Miss. value count	Action	Replace value
BOD	2	Replaced	8
COD	1	Replaced	26
SS	3	Replaced	94
TUR	15	Replaced	80.900
E. coli	10	Replaced	99596
Coliform	10	Replaced	98498046
NO <sub>3</sub>	3	Replaced	2.570
Hg	5	Replaced	0.0005
PO <sub>4</sub>	36	Replaced	0.150

values' magnitudes will scale to appreciably low values [48, 49]. The attributes in this dataset were normalized using the Unsupervised Attribute Normalize filter, which rescales each numeric attribute to a range of 0 to 1. Table 6 shows the detailed data.



TABLE 3: Replace to round number.

Attribute	Original value	Replace value	Count
SS	<1.000	1.000	1
NH <sub>3</sub> -NL	<0.010	0.010	16
<i>E. coli</i>	<1.000	1.000	16
NO <sub>3</sub>	<0.010	0.010	55
Cl	<1.000	1.000	6
PO <sub>4</sub>	<0.010	0.010	93
As	<0.001	0.001	17
Hg	<0.001	0.001	815
Cd	<0.001	0.001	865
Cr	<0.001	0.001	677
Pb	<0.010	0.010	674
Zn	<0.010	0.010	83
Ca	<0.100	0.100	10
Fe	<0.010	0.010	63
K	<0.100	0.100	1
Mg	<0.100	0.100	39

**2.3. Machine Learning Task.** In this process, four classifications and three feature selection algorithms were used. The classification algorithms were ANN, SVM, Random Forest, and Naïve Bayes. Meanwhile, the feature selection algorithms were the Information Gain Attribute Evaluation, Correlation Attribute Evaluation, and Symmetrical Uncertainty Attribute.

**2.3.1. Classification Algorithm.** An ANN is a series of interconnected input-output networks, each with its weight. This system consists of one input layer, one or more intermediate layers, and one output layer. In this study, the weight of relation was adjusted to train a neural network. The network's efficiency was improved by iteratively modifying the weight [50]. When it comes to pattern recognition, this model excelled. Although the model can quickly adjust to new data values, it can slowly converge and risk a local optimum [51].

Meanwhile, SVM is a well-known machine learning technique, primarily for classification. Theoretically, SVM is defined as a discriminative classifier with an optimal hyperplane. The optimal hyperplane generates support vectors, which classify new examples and datasets that support the hyperplane. This hyperplane is a line that divides the two-dimensional (2D) region into two segments, with each element lying on either side. For example, multiple line data classification was completed with two distinct datasets (squares and dots) and suggested an affirmative interpretation. However, selecting the best hyperplane is challenging because it must be noise-free and accurately generalize datasets. SVM is attempting to discover an efficient hyperplane that provides a significant minimum distance to the qualified data collection point [52, 53].

On the other hand, Random Forest is a data mining technique for solving classification and regression issues. Growing an ensemble of trees and voting to determine the class category greatly improved classification accuracy. Random vectors are created to develop these ensembles. Each tree is constructed

from one or more random vectors. Random Forest is also made up of classifier and regression trees. The production of trees is analyzed to solve classification problems. Most class votes determine the Random Forest prediction. Since overfitting does not occur in random sizeable forests, the generalization error merges to a limiting value as more trees are added to the Random Forest [54].

Next, the Naïve Bayes algorithm is a simple probabilistic classifier that computes a set of probabilities by counting the frequency and combinations of values in a dataset. When calculating the value of the class variable, the algorithm employs Bayes' theorem and assumes that all variables are independent. Although this conditional independence assumption is rarely relevant to real-world applications, the algorithm could perform in various supervised classification problems [51, 55].

**2.3.2. Feature Selection.** The feature selection process excludes features that do not play a significant role in determining water quality status. It substantially affects data dimension measurement, whether for data training or data testing. The four approaches to select the feature subset, in general, are filters (a process of feature evaluation conducted independently from the learning process), wrappers (a method of feature subset selection based on the learning process evaluation result), embedded (a feature selection conducted during the learning process), and simple filters (this approach is usually used on data with many features such as on the case of textual classification).

This study employed a feature selection method to distinguish the best feature subset from the learning process. Each feature subset's score or weight was used to determine the best subset [6]. The stage of the filter approach is shown in Figure 2. This study employed three algorithms from the filters approach category: information gain, correlation, and symmetrical uncertainty to quantify the weight value.

**(1) Feature Selection Algorithm.** Information Gain Attribute Evaluation calculates each attribute's information gain or entropy in the dataset class. The values range from 0 to 1, with 0 indicating "no information" and 1 representing "maximum information." The benefits feature was applied to the current attributes to distinguish between higher and lower information gains. The correlation between two variables was calculated using the Correlation Attribute Evaluation [56], which can be in the same or opposite directions on the number line. The determined correlation between each attribute and the output variable would reveal which attributes rank higher from a moderate-to-high positive or negative correlation (close to -1 or 1). Attributes with low correlation (values close to zero) can be removed from the selection.

The information gain's bias against features with more values was compensated by symmetrical uncertainty, which normalizes its values within the range [0, 1], with 1 indicating that knowledge of one of the values fully predicts the value of the other and 0 indicating that  $X$  and  $Y$  are independent. It takes an asymmetrical approach for a pair of functions. The symmetrical uncertainty value serves two purposes: (1) it can delete features with symmetrical uncertainty less than the threshold, and (2) it can calculate the weight of each function,

TABLE 4: Description of the attributes used for classification.

	Attribute	Possible values			
		Min	Max	Mean	Std dev
DOE-WQI	DO	0.870	11.140	6.010	1.618
	BOD	1	47	8.090	5.410
	COD	3	167	26.160	17.215
	SS	1	2690	93.760	188.745
	PH	3.800	9.436	7.192	0.539
	NH3-NL	0.010	14.600	1.815	2.240
National WQS	COND	4	38412	1021.721	4496.949
	SAL	0	24.300	0.591	2.751
	DS	16	24300	665.11	2947.565
	TEMP	23.260	34.300	28.690	1.787
	TUR	0	1262.428	80.890	111.089
	E. coli	0	$5.800 \times 10^7$	99596.149	1928195.220
	Coliform	0	$4.700 \times 10^{11}$	98488046.287	1673957603.307
All chemical contents	TS	23	24338	755.442	2956.568
	NO3	0.010	126	2.57	7.208
	Cl	0.965	13500	288.784	1506.680
	PO4	0.010	3.510	0.151	0.242
	As	0.001	0.108	0.007	0.007
	Hg	0	0.006	0	0.001
	Cd	0.001	0.006	0.001	0
	Cr	0.001	0.049	0.001	0.002
	Pb	0	0.195	0.009	0.010
	Zn	0.005	0.834	0.045	0.055
	Ca	0	3160	18.049	109.088
	Fe	0.010	3.150	0.457	0.389
	K	0.100	500	11.354	38.312
	Mg	0.100	2600	20.906	126.100
	Na	0.114	6750	162.623	816.219

TABLE 5: Outliers and extreme values.

	Label	DOE-WQI Attribute (7) instance (913) Count	National Water Quality Standard Attribute (14) instance (913) Count	All chemical content Attribute (28) instance (913) Count
Outliers	No	849	771	674
	Yes	64	142	239
Extreme values	No	896	775	640
	Yes	17	138	273

which will be used to direct population initialization for genetic algorithms in the memetic system. The function with the highest symmetrical uncertainty value would receive more weight.

### 3. Results and Discussion

The testing procedure used in this study employed  $k$ -fold cross-validation, which divides the dataset into two parts,

one for training and the other for testing, and then alternates the two parts ten times. We also applied a split test in which the data was divided into 70:30 ratios, with 70% randomly sampled for model training and 30% for independent research. Accuracy, precision, and recall were some of the parameters used to compare the performance of one model to another [57]. This research also used a statistical test to determine whether one classifier's performance is statistically different from another.

TABLE 6: Description of the attributes used for classification.

	Attribute (7) Instance (833)	Possible value			
		Min	Max	Mean	Std dev
I DOE-WQI	DO	0	1	0.505	0.156
	BOD	0	1	0.272	0.180
	COD	0	1	0.254	0.162
	SS	0	1	0.196	0.203
	PH	0	1	0.491	0.127
	NH3-NL	0	1	0.187	0.209
	Attribute (13) Instance (619)	Possible value			
		Min	Max	Mean	Std dev
National Water Quality Standard	DO	0	1	0.522	0.152
	BOD	0	1	0.261	0.175
	COD	0	1	0.253	0.163
	SS	0	1	0.185	0.192
	PH	0	1	0.495	0.124
	NH3-NL	0	1	0.179	0.210
	COND	0	1	0.259	0.177
	SAL	0	1	0.225	0.176
	DS	0	1	0.214	0.191
	TEMP	0	1	0.484	0.155
	TUR	0	1	0.209	0.188
	E-coli	0	1	0.116	0.174
	Coliform	0	1	0.157	0.196
	Attribute (28) Instance (533)	Possible value			
		Min	Max	Mean	Std dev
All chemical contents	DO	0	1	0.538	0.149
	BOD	0	1	0.244	0.171
	COD	0	1	0.236	0.161
	SS	0	1	0.170	0.181
	PH	0	1	0.495	0.122
	NH3-NL	0	1	0.156	0.193
	COND	0	1	0.281	0.178
	SAL	0	1	0.240	0.180
	DS	0	1	0.204	0.177
	TS	0	1	0.173	0.130
	TEMP	0	1	0.472	0.157
	TUR	0	1	0.205	0.180
	E. coli	0	1	0.115	0.174
	Coliform	0	1	0.158	0.197
	NO3	0	1	0.210	0.201
	Cl	0	1	0.217	0.172
	PO4	0	1	0.186	0.198
	As	0	1	0.168	0.138
	Hg	0	1	0.412	0.414
	Cd	0	1	0	0
	Cr	0	1	0	0
	Pb	0	1	0.592	0.341
	Zn	0	1	0.191	0.182
	Ca	0	1	0.264	0.223



TABLE 6: Continued.

Fe	0	1	0.271	0.204
K	0	1	0.314	0.158
Mg	0	1	0.216	0.169
Na	0	1	0.172	0.141

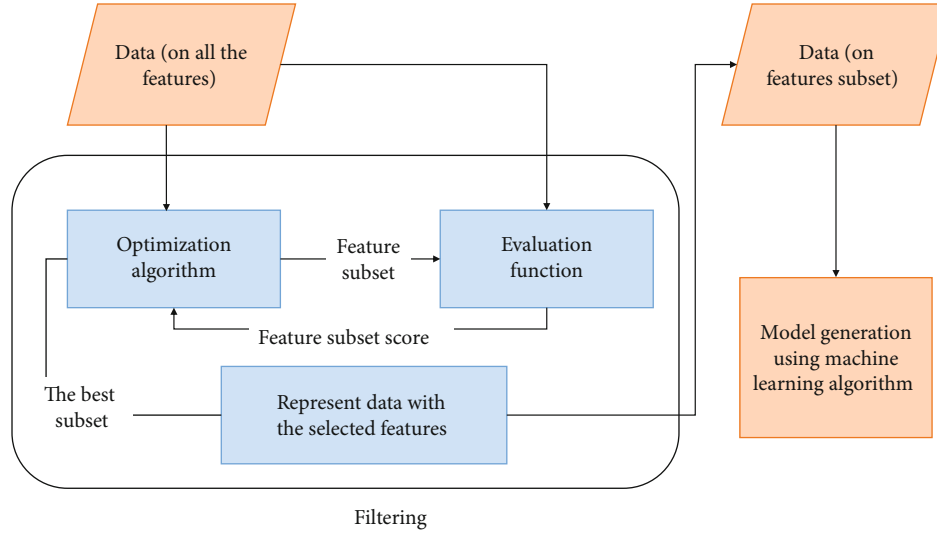


FIGURE 2: The stage of the process of feature subset finding using filters approach [6].

A comparative experiment is usually divided into stages. First, the feature selection algorithms are compared to determine which attributes can be omitted as irrelevant and essential for predicting water quality. The info gain, correlation, and symmetrical uncertainty will be used to test feature selection algorithms on the data collection. The results of contrasting feature selection algorithms on the dataset are shown in Table 7. Table 7 demonstrates the same top five attributes in the DOE-WQI dataset for the correlation, information gain, and symmetrical uncertainty (DO, NH<sub>3</sub>-NL, BOD, COD, and SS). Meanwhile, the eight most attributes for the National Water Quality Standard dataset were DO, COND, SAL, NH<sub>3</sub>-NL, DS, BOD, COD, and TUR. And for all chemical contents, the top 10 investigated were COND, SAL, DO, DS, TS, NH<sub>3</sub>-NL, Na, Cl, BOD, and COD.

Then, as shown in Table 7, new datasets are created after analyzing the results and removing the unnecessary attributes that we found. After that, Random Forest was selected as the initial classifier to be compared against two test options, cross-validation and percentage split with the new datasets of normalized correlation, information gain, and symmetrical uncertainty which generated after employing feature selection algorithm. We choose information gain feature selection as it achieved the highest accuracy for cross-validation test option and also percentage split for overall new selected datasets.

Subsequently, we applied the sampling method to improve the performance of the classification accuracy. The imbalance data using an adaptive oversampling technique was detected using SMOTE. Consequently, the minority class's prediction

accuracy will be improved. Figure 3 shows a bar graph of SMOTE application in the dataset at 400% oversampling degree with five nearest neighborhoods.

The results were then compared to three other classifiers in the following experiment to determine the best classification classifier. In this phase, the ANN classifier was compared to SVM, Random Forest, and Naïve Bayes for two test options and three datasets chosen in the first experiment. Figures 4(a)–4(f) compare the accuracy efficiency, recall, precisions of each classifier, and the result of the SMOTE process before and after. The result showed that the normalized dataset using Random Forests had higher accuracy using split test 70%–30% in DOE-WQI with six attributes data with an accuracy of 95.64% and recall and precision close to 1.

However, only minor changes were observed when we applied the SMOTE to the data. Hence, in practice, the data is almost imbalanced between the classes. Consequently, we chose not to use SMOTE in this study as it would change the actual data situation and had no significant impact on increasing the data accuracy. Moreover, Figure 4(a) consists of all six attributes of the DOE-WQI dataset, whereby Figure 4(b) consists of five attributes. Nonetheless, both data showed the highest accuracy when Random Forest was used as a classifier.

**3.1. Statistical Test.** Several statistical tests were run to validate the Random Forest classifier results. A paired corrected *t*-test was used to compare output between learning methods for both the cross-validated and percentage-split datasets. This test compared paired samples to determine whether

TABLE 7: Top attribute comparison of feature selection algorithms.

Feature selection	Correlation			Information gain			Symmetrical uncertainty		
	Average merit	Rank	Att.	Average merit	Rank	Att.	Average merit	Rank	Att.
IK DOE-WQI	$0.494 \pm 0.006$	1	DO	$0.481 \pm 0.012$	1	DO	$0.275 \pm 0.005$	1	NH3-NL
	$0.441 \pm 0.006$	2	NH3-NL	$0.452 \pm 0.011$	2	NH3-NL	$0.275 \pm 0.008$	2	DO
	$0.377 \pm 0.007$	3	BOD	$0.450 \pm 0.011$	3	BOD	$0.248 \pm 0.006$	3	BOD
	$0.369 \pm 0.008$	4	COD	$0.414 \pm 0.009$	4	COD	$0.238 \pm 0.005$	4	COD
	$0.251 \pm 0.010$	5	SS	$0.151 \pm 0.010$	5	SS	$0.115 \pm 0.007$	5	SS
National Water Quality	$0.520 \pm 0.007$	1	DO	$0.519 \pm 0.010$	1	COND	$0.317 \pm 0.010$	1	COND
	$0.503 \pm 0.007$	2	COND	$0.512 \pm 0.012$	2	NH3-NL	$0.307 \pm 0.008$	2	DS
	$0.502 \pm 0.006$	3	SAL	$0.507 \pm 0.013$	3	DS	$0.305 \pm 0.007$	3	NH3-NL
	$0.490 \pm 0.002$	4	NH3-NL	$0.505 \pm 0.011$	4	SAL	$0.296 \pm 0.012$	4	SAL
	$0.477 \pm 0.006$	5	DS	$0.477 \pm 0.012$	5	DO	$0.280 \pm 0.009$	5	DO
	$0.399 \pm 0.006$	6	BOD	$0.440 \pm 0.015$	6	BOD	$0.243 \pm 0.008$	6	COD
	$0.388 \pm 0.006$	7	COD	$0.415 \pm 0.006$	7	COD	$0.239 \pm 0.008$	7	BOD
	$0.280 \pm 0.014$	8	TUR	$0.196 \pm 0.013$	8	TUR	$0.134 \pm 0.007$	8	TUR
All chemical contents	$0.553 \pm 0.007$	1	COND	$0.569 \pm 0.014$	1	COND	$0.245 \pm 0.013$	1	COND
	$0.550 \pm 0.007$	2	SAL	$0.529 \pm 0.009$	2	NH3-NL	$0.315 \pm 0.007$	2	NH3-NL
	$0.537 \pm 0.008$	3	DO	$0.529 \pm 0.016$	3	SAL	$0.314 \pm 0.008$	3	DS
	$0.503 \pm 0.008$	4	DS	$0.525 \pm 0.015$	4	DS	$0.306 \pm 0.009$	4	Na
	$0.497 \pm 0.007$	5	TS	$0.487 \pm 0.015$	5	TS	$0.303 \pm 0.012$	5	SAL
	$0.489 \pm 0.005$	6	NH3-NL	$0.473 \pm 0.010$	6	DO	$0.300 \pm 0.016$	6	TS
	$0.461 \pm 0.006$	7	Na	$0.459 \pm 0.012$	7	BOD	$0.285 \pm 0.007$	7	DO
	$0.444 \pm 0.008$	8	Cl	$0.451 \pm 0.014$	8	Na	$0.258 \pm 0.007$	8	COD
	$0.435 \pm 0.009$	9	K	$0.429 \pm 0.012$	9	COD	$0.258 \pm 0.004$	9	BOD
	$0.422 \pm 0.008$	10	Ca	$0.358 \pm 0.020$	10	K	$0.233 \pm 0.007$	10	K

two learning models were statistically significantly different or more significant to one other. The ANN, SVM, Random Forest, and Naïve Bayes classifiers were tested on the DOE-WQI, National Water Standard, and all chemical content datasets with 0.05 twin-tailed confidence.

From Figure 5, a cross-validation  $t$ -test was performed on the DOE-WQI dataset with (a) 6 attributes and (b) 5 attributes. We compared those attributes to identify whether they were significant. This study only compared DOE-WQI data, providing the highest accuracy from the previous test. Naïve Bayes was selected as the base learning model over the other three in both selections. Based on the test in (a), ANN and SVM were statistically different but not better than the Naïve Bayes. When ANN became a base, Naïve Bayes and SVM were significantly different. When SVM became a base, another classifier produced the same result, but Random Forest was still a substantially better base. Instead, the Random Forest classifier was proven to be significantly better than the base Naïve Bayes classifier.

Meanwhile, in (b), when Naïve Bayes acted as a base, SVM differed substantially, while ANN and Random Forest performed significantly better. In contrast, when ANN was set as the base, Naïve Bayes was markedly lower, SVM was different, and Random Forest was better. However, when

SVM became the base, the result revealed that Naïve Bayes and ANN were significantly different. Nonetheless, Random Forest was substantially better than the base.

Subsequently, the percentage-split  $t$ -test was run on the DOE-WQI dataset as (a) and (b). All Random Forest, ANN, and SVM models in (a) differed significantly from the base Naïve Bayes model. When ANN was used as a base, the results were similar to Naïve Bayes but quite different. After which, when SVM was used as the base classifier, Naïve Bayes and ANN became significantly different. In comparison, the Random Forest classifier was substantially better than the SVM base. In (b), ANN and SVM were markedly different from the base, but Random Forest performed better than Naïve Bayes. Nevertheless, when ANN became the base, all others differed significantly. With SVM as the base, Naïve Bayes and ANN remained different, but Random Forest outperformed the base.

Based on this analysis, Random Forest has always been significantly better than the base. We can compare (a) and (b) with different attributes, showing that they were not significant with others. In cross-validation, the results showed that (a) was significantly better than the base in percentage split  $t$ -test, while (b) was significantly higher.

The cross-validation  $t$ -test on the National Water Quality Standard is shown in Figure 6. First, when Naïve Bayes

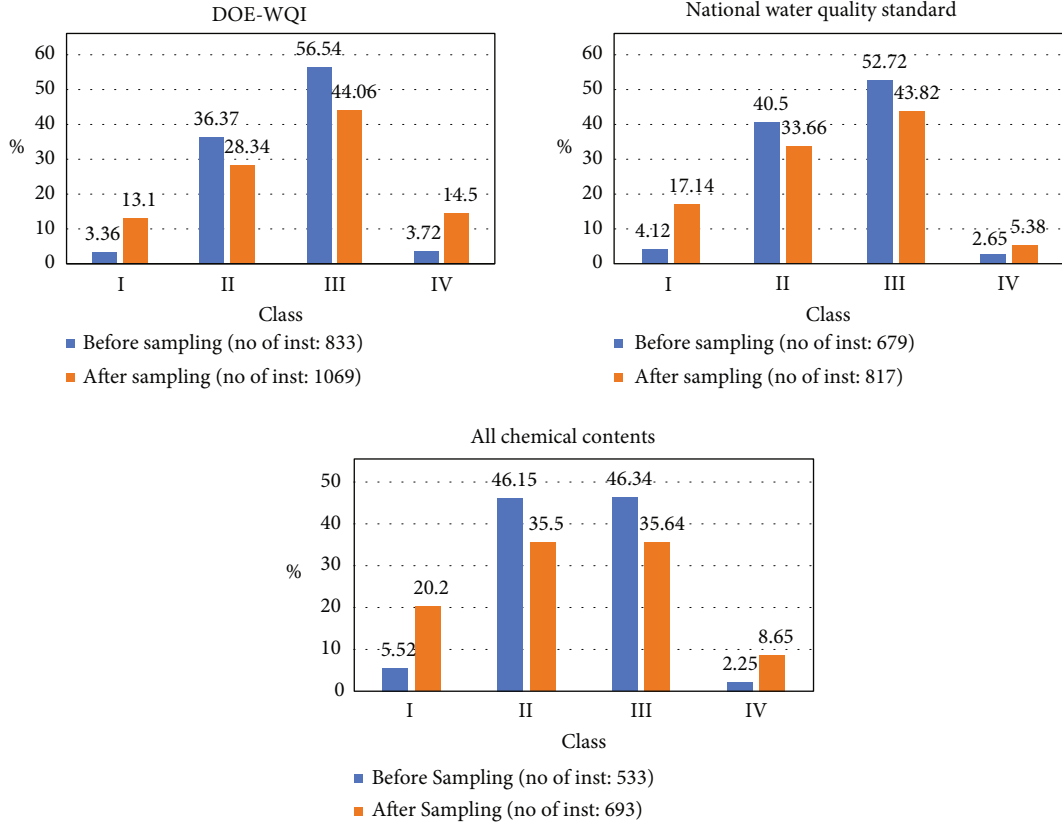


FIGURE 3: The dataset after SMOTE was applied to DOE-WQI, National Water Quality Standard, and all chemical contents.

was chosen as a base, all three classifiers performed significantly better. When ANN became a base, Naïve Bayes was substantially lower. Meanwhile, SVM and Random Forest were distinct from the base. As SVM became the base, Naïve Bayes was significantly lower, ANN was very different, and Random Forest was better. Even so, Random Forest remained the best classifier when compared to others. This finding was also proved in the percentage-split  $t$ -test performed on National Water Quality Standard, where Random Forest served significantly better than ANN, SVM, and Naïve Bayes.

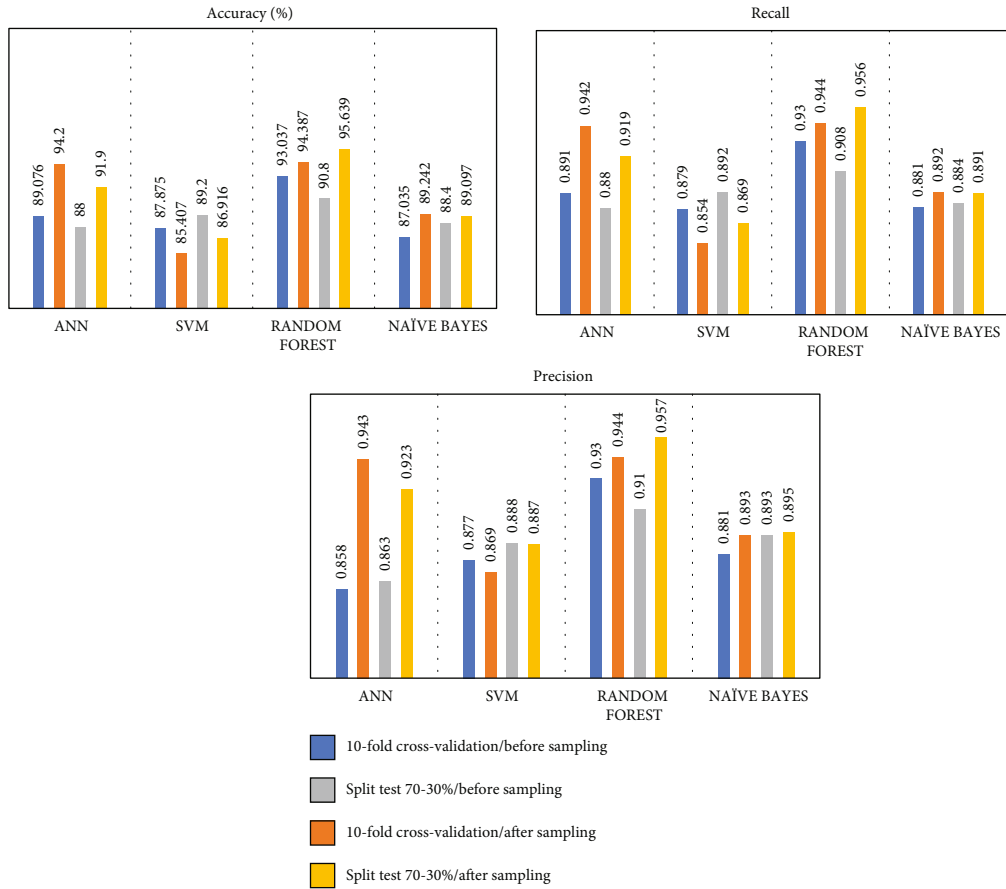
The all chemical content data were subjected to a cross-validation  $t$ -test in Figure 7. When Naïve Bayes became a base, the other classifiers performed significantly better than the base. However, Naïve Bayes became considerably lower when ANN was used as a base, and the SVM and Random Forest became significantly different. Next, as SVM worked as a base, and Naïve Bayes demonstrated that it was significantly lower, ANN became wildly dissimilar to the base, and Random Forest was considerably better.

Meanwhile, when Naïve Bayes acted as a base in the percentage-split  $t$ -test, the results were the same; all three classifiers were significantly better. As ANN became a base, the performance of Naïve Bayes and SVM decreased, but Random Forest remained incomparable. Finally, when SVM was used as a base, the performance of Naïve Bayes dropped, but ANN and Random Forest performed significantly better than the base. The statistical test conducted showed that the Random Forest outperformed the classifier algorithm, with a statistically significant difference in the outcomes.

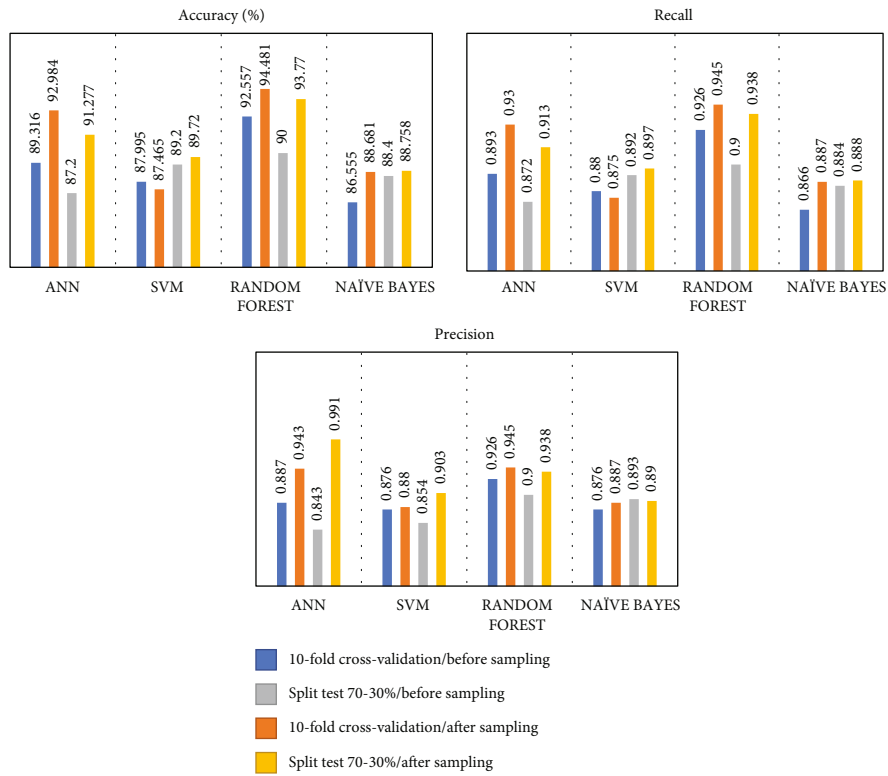
**3.2. Regularization in Machine Learning.** In machine learning, the principle of regularization is used to avoid overfitting for the learning model constructed from training data. It is a parameter in the classification problem that is used to minimize the error function. A few strategies were used in this study to resolve and restrict the issue of overfitting. The techniques mainly involved using feature selection algorithms to reduce the number of features used in this analysis, cross-validation during testing, and an oversampling approach with SMOTE to generate more data for the training set.

**3.3. Decision Tree Classifier.** In the decision tree-based feature selection approach, constructing a decision tree is referred to as feature selection. Each layer's partition samples feature is calculated according to standards, and the essential feature is consistently selected as the feature of partition samples. The feature selection method based on a decision tree includes constructing a decision tree as part of the feature selection process. When determining each layer's features of partition samples, each feature is calculated according to specific standards, and the most crucial feature is frequently selected [58].

The decision tree algorithm's key advantages are its high classification accuracy and robustness. The issue is that it is prone to overfitting, that samples immediately impact decision trees, and that the subtree may appear multiple times. Pruning technology and  $k$ -fold cross-validation can be used to combat overfitting. Unnecessary branches can be pruned ahead of time; hence, overfitting is possible to avoid. For the second issue, a prefiltering phase in the preprocessing

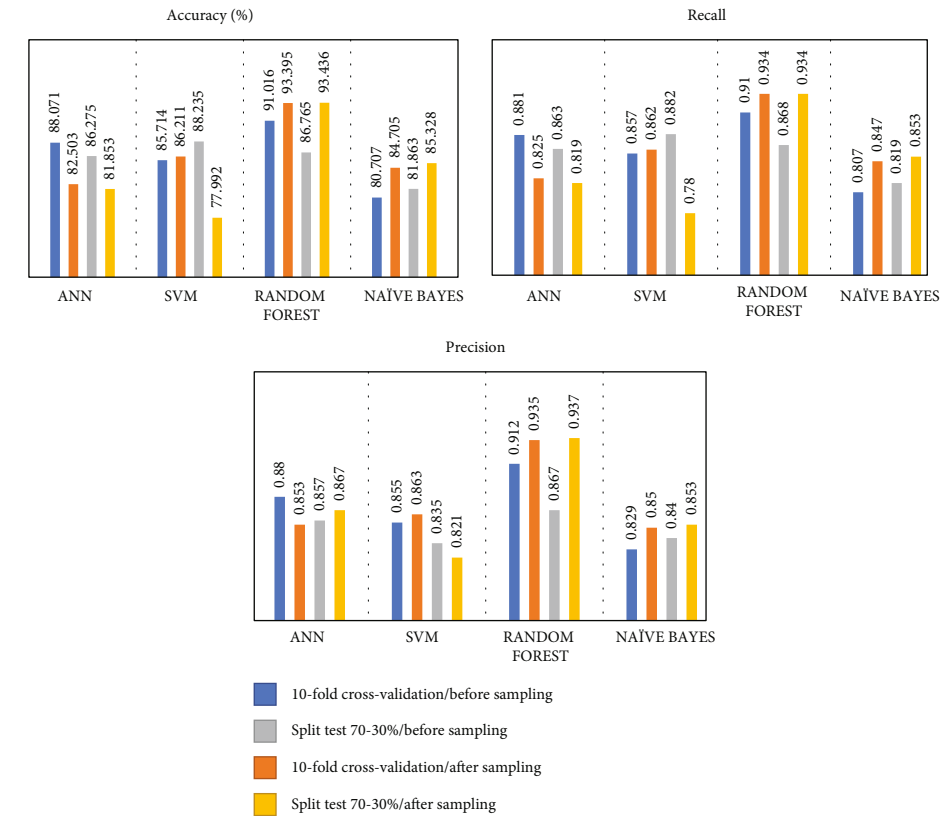


(a)

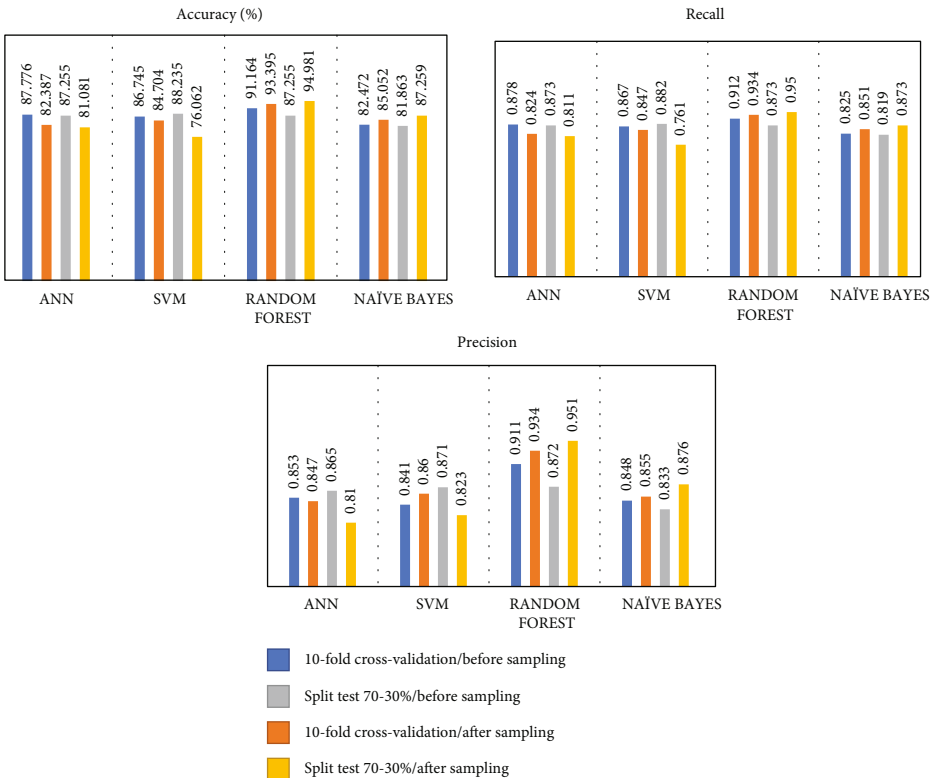


(b)

FIGURE 4: Continued.



(c)



(d)

FIGURE 4: Continued.



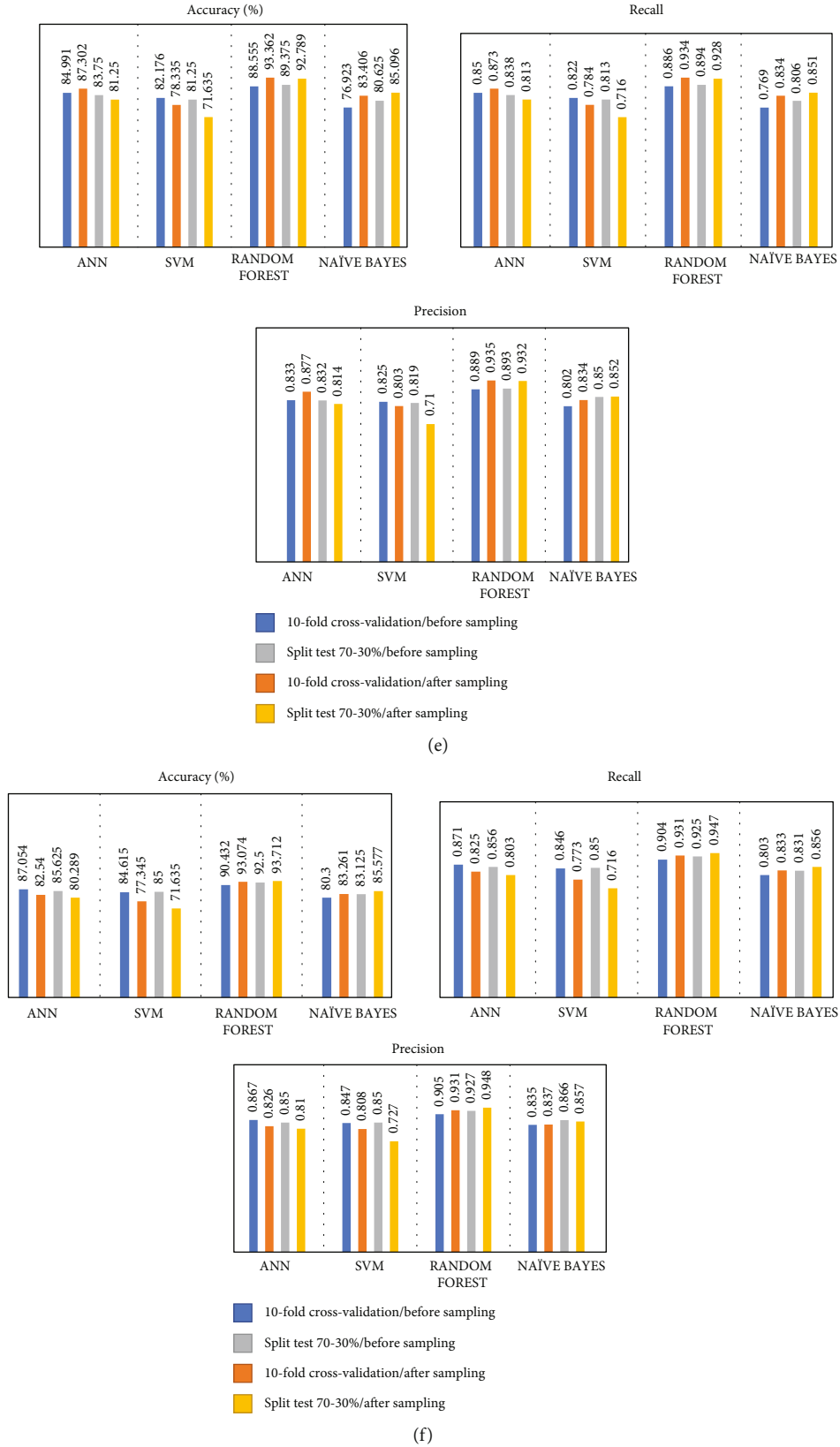


















































FIGURE 4: (a) DOE-WQI (all attributes); (b) DOE-WQI (5 attributes); (c) National Water Quality Standard (all attributes); (d) National Water Quality Standard (8 attributes); (e) all chemical contents (all attributes); (f) all chemical contents (10 attributes).

Cross-validation split T-Test result			6 Attributes			Percentage split (70-30) T-Test Result		
			Paired corrected T-Test					
			Percentage of correctly classified					
			Normalize DOE-WQI					
Confidence			0.05(two-tailed)					
Classifier			Result					
Naïve Bayes	(86.190)	(86.190)	(86.190)		(85.800)	(85.800)	(85.800)	
								
ANN	(88.090)	(88.090)	(88.090)		(88.630)	(88.630)	(88.630)	
								
SVM	(87.540)	(87.540)	(87.540)		(87.540)	(87.540)	(87.540)	
								
Random forest	(91.580 v)	(91.580 v)	(91.580 v)		(91.250)	(91.250)	(91.250 v)	
								

(a)

Cross-validation split T-Test result			Percentage split (70-30) T-Test result			
Paired corrected T-Test Percentage of correctly classified Normalize DOE-WQI 0.05(two-tailed)						
Confidence	Result					
Classifier						
Naïve Bayes	(86.090) 	(86.090*) 	(86.090) 	(86.640) 	(86.640) 	(86.640) 
ANN	(89.090 v) 	(89.090) 	(89.090) 	(88.820) 	(88.820) 	(88.820) 
SVM	(87.670) 	(87.670) 	(87.670) 	(87.360) 	(87.360) 	(87.36) 
Random forest	(91.850 v) 	(91.850 v) 	(91.850 v) 	(91.850 v) 	(91.850) 	(91.850 v) 



(b)

FIGURE 5: DOE-WQI.





























	Cross-validation split T-Test result			Percentage split (70-30) T-Test result		
	Paired corrected T-Test					
	Percentage of correctly classified					
	Normalize national water quality standard					
Confidence	0.05 (two-tailed)					
Classifier	Result					
Naïve bayes	(80.710) 	(80.710*) 	(80.710*) 	(81.770) 	(81.770*) 	(81.770*) 
ANN	(88.080 v) 	(88.080) 	(88.080) 	(87.190 v) 	(87.190) 	(87.190) 
SVM	(85.720 v) 	(85.720) 	(85.720) 	(88.180 v) 	(88.180 v) 	(88.180) 
Random forest	(91.020 v) 	(91.020) 	(91.020 v) 	(89.660 v) 	(89.660 v) 	(89.660 v) 
<div><div> Base Comparison</div><div> Significantly better than the base</div><div> Significantly different to base</div><div> Significantly lower than the base</div></div>						



FIGURE 6: National Water Quality Standard.

Confidence Classifier	Cross-validation split T-Test result		Percentage split (70-30) T-Test result			
	Paired corrected T-Test		Percentage of correctly classified			
	Normalize All chemical content		0.05 (two-tailed)			
	Result		Result			
Naïve Bayes	(76.930)	(76.930*)	(76.930*)	(80.750)	(80.750*)	(80.750*)
ANN	(84.990 v)	(84.990)	(84.990)	(84.470 v)	(84.470)	(84.470 v)
SVM	(82.180 v)	(82.180)	(82.180)	(83.230 v)	(83.230*)	(83.230)
Random forest	(88.150 v)	(88.150)	(88.150 v)	(90.060 v)	(90.060 v)	(90.060 v)

● Base Comparison  
 ● Significantly better than the base  
 ● Significantly different to base  
 ● Significantly lower than the base

FIGURE 7: All chemical content.

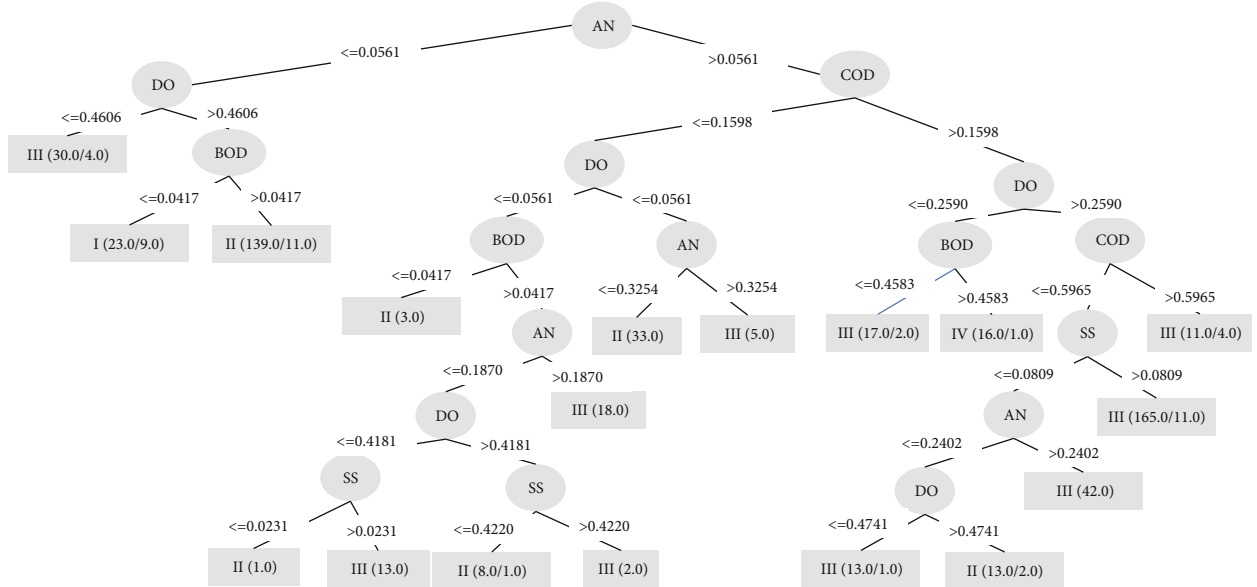


FIGURE 8: The decision tree from 6 attributes with reducing error pruning.

data stage could eliminate some irrelevant features. As a result, the decision tree's size can be reduced, and erroneous decision trees could be avoided [59].

We utilized the feature selection algorithm to filter the features before constructing the decision tree, removing those with low correlation with the category and keeping those with high correlation as the feature subset of the next stage of decision tree construction. In this work, we generated a pruned decision tree from cross-validation data 10-fold before recorded the minimum object one at which the attribute was tested in the tree to determine how dependent one attribute is on others. The decision tree will show all the possible scenarios implied by the dependency diagram. Decision trees can be pruned to reduce their size and improve their interpretability. Based on the previous results, Random Forest had emerged as one of the most accurate

classifiers. Therefore, in this section, we tried to construct a tree representation that may be used to assess the importance of features. In this experiment, we used the tree classifier J48 to visualize the tree. It is a dependable, robust, and simple-to-understand tree-generation method for data mining.

Figure 8 depicts that at the top level of this tree, the root node, DO, gives the highest number of weights (0.46), which was vital in this experiment when ammonia was lower than or equal to 0.06. While selecting the DO, the left branch corresponding to the possible values of DO should be considered and planned. There was no more splitting because the other branch had already reached the quest for purity. This study is deemed to be completed after it met a pure node.

Meanwhile, Figure 9 shows the root node at the top level of the decision tree. We can see that BOD gave the most weight

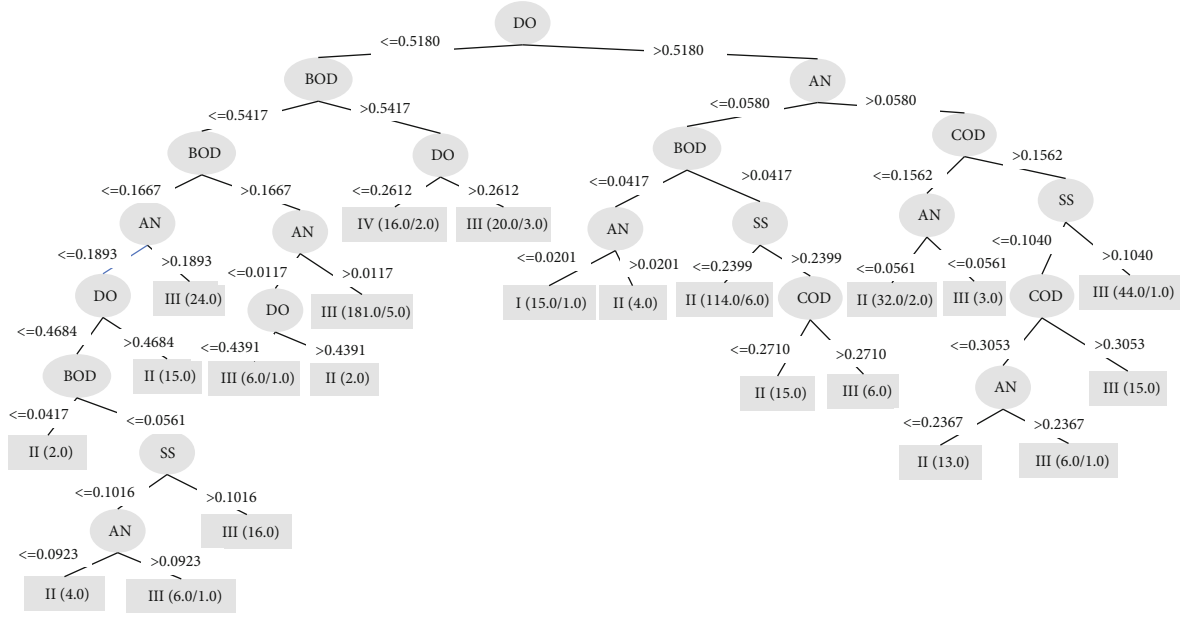


FIGURE 9: This The decision tree from 5 attributes with reducing error pruning.

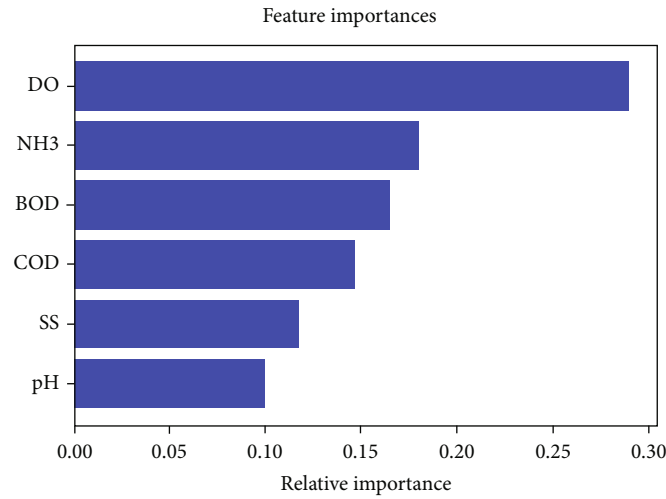


FIGURE 10: Important feature.

(0.54), which was the most critical feature in this experiment when DO was lower than or equal to 0.52. BOD will be selected in this case, and the next branch with a higher weight's value was DO (0.26 weights). Then, the end of the node did not split as it reached the quest for purity. The decision tree might be easily influenced by the category and irrelevant features. Creating the decision tree in this scenario is complicated, and the likelihood of overfitting is high. Therefore, the most relevant characteristic must be chosen when constructing the decision tree.

**3.4. SHapley Additive exPlanations (SHAP) on Water Quality Features.** Additionally, we employed a unified model explanation system, SHapley Additive exPlanations (SHAP), to ensure

the consistency of the explanation technique for various models and to strive for a more thorough and informative interpretation [60, 61]. The contribution of each feature to the model prediction is displayed using the SHAP values. The impact of a feature on the model output increases with the absolute SHAP value of the feature; positive values increase output, while negative values decrease output. Every symbol's color (red for high; blue for low) denotes the importance of the features in the input dataset [62]. We created SHAP summary graphs based on the SHAP values for the RF trained model in order to analyze and validate the analytic results from a different perspective. The RF models developed for this study were capable of accuracy comparable to other models utilized in earlier studies. The best models' feature

importance based on SHAP values were determined to identify the key contributors to the water quality as shown in Figure 10. The most important feature shown in the figure was DO with relative importance 0.29.

#### 4. Conclusions

It can be concluded that the feature selection algorithms applied were found to classify the most convenient features. The selected features could solve the classification problem of predicting water quality relevant to this study and the available dataset. The DOE-WQI dataset included the DO, NH<sub>3</sub>-NL, BOD, COD, SS, and pH. The efficacy of the Random Forest classifier outperformed Naïve Bayes, ANN, and SVM with statistically significant differences in the three classification algorithms' outputs. Our findings discovered only slight differences in accuracy when applying DOE-WQI tests with all six attributes. As we attempted to reduce it to five attributes, we found that the outcomes were acceptable with high accuracy. As mentioned earlier in this study, the parameters' number can affect processing time and cost. Thus, after constructing a survey with a decision tree and selecting a feature, we found that DO and BOD were the most relevant features, demonstrating that both were the primary components in predicting and determining water quality. Therefore, our study can conclude that, due to the algorithm's success with feature selection, even with a minor water quality feature from the dataset, better accuracy is feasible to achieve.

We also take a look at previous research studies that have studied important features in determining water quality. The study examined two different models, namely feed-forward back propagation neural network (BPNN) and radial basis function neural network (RBFNN), which have been done before. Still, it only focused on features in WQI which found that DO influence the WQI prediction by higher, followed by COD, SS, NH<sub>3</sub>-N, and pH, respectively [63]. Next, the classification of WQI within a particular water quality class was predicted using a decision tree machine learning technique studied [64]. The findings indicate that the metrics with the highest association to WQI prediction are the BOD, COD, and DO. Another study is to develop an input method employing ANNs to compute the WQI from input parameters rather than using the parameter indices when one of the parameters is missing. The sensitivity analysis revealed that DO has the greatest impact on WQI [65]. For this research, the limitation of the study lies in the fact that the results obtained from this study are specific to the region, and thus, consideration should be taken when extending the results to other regions and other water quality parameters with wide variations. In the future, an attempt will be made to visualize the results using augmented reality for better monitoring. The data will be displayed in three-dimension (3D), and an interaction technique will be implemented to improve the data's interaction.

#### Data Availability

The data used for the findings of this study is available upon request from the corresponding author.

#### Conflicts of Interest

The authors declare they have no conflicts of interest to report regarding the present study.

#### Acknowledgments

This research was funded by the Universiti Kebangsaan Malaysia (Grant code: GUP2019-060). This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (Ministry of Science and ICT, South Korea) (No. 2022-0-01200, Training Key Talents in Industrial Convergence Security).

#### References

- [1] M. R. Golabi, S. Farzi, F. Khodabakhshi, F. S. Geshnigani, F. Nazdane, and F. Radmanesh, "Biochemical oxygen demand prediction: development of hybrid wavelet-random forest and M5 model tree approach using feature selection algorithms," *Environmental Science and Pollution Research*, vol. 27, no. 27, pp. 34322–34336, 2020.
- [2] S. N. Jamali, M. S. Zakaria, and S. N. Jamali, "A risk management approach to the development of an early warning system: a case for Tasik Chini," *APJITM*, vol. 7, no. 2, pp. 115–130, 2018.
- [3] H. Effendi, "River water quality preliminary rapid assessment using pollution index," *Procedia Environmental Sciences*, vol. 33, pp. 562–567, 2016.
- [4] I. Lee, H. Hwang, J. Lee, N. Yu, J. Yun, and H. Kim, "Modeling approach to evaluation of environmental impacts on river water quality: a case study with Galing River, Kuantan, Pahang, Malaysia," *Ecological Modelling*, vol. 353, pp. 167–173, 2017.
- [5] E. Salahat and M. Qasimeh, "Recent advances in features extraction and description algorithms: a comprehensive survey," in *2017 IEEE international conference on industrial technology (ICIT)*, Toronto, ON, Canada, 2017.
- [6] S. Uyun and E. Sulistyowati, "Feature selection for multiple water quality status: integrated bootstrapping and SMOTE approach in imbalance classes," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, p. 4331, 2020.
- [7] M. F. Kasim, H. Juahir, I. Tawnie et al., "Environmetric techniques application in water quality assessment: a case study in Linggi River basin," *Jurnal Teknologi*, vol. 74, no. 1, 2015.
- [8] S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, "Water quality assessment in terms of water quality index," *American Journal of water resources*, vol. 1, no. 3, pp. 34–38, 2013.
- [9] A. Abu, R. Hamdan, and N. S. Sani, "Ensemble learning for multidimensional poverty classification," *Sains Malaysiana*, vol. 49, no. 2, pp. 447–459, 2020.
- [10] R. Alhutaish and N. Omar, "Feature selection for multi-label document based on wrapper approach through class association rules," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 2, pp. 642–649, 2017.
- [11] T. Yu and Y. Bai, "Comparative study of optimization intelligent models in wastewater quality prediction," in *2018 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, Xi'an, China, 2018.



- [12] Y. Cao, Y. Ye, H. Zhao et al., "Remote sensing of water quality based on HJ-1A HSI imagery with modified discrete binary particle swarm optimization-partial least squares (MDBPSO-PLS) in inland waters: a case in Weishan Lake," *Ecological Informatics*, vol. 44, pp. 21–32, 2018.
- [13] R. Xu, Q. Xiong, H. Yi, C. Wu, and J. Ye, "Research on water quality prediction based on SARIMA-LSTM: a case study of Beilun Estuary," in *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, Zhangjiajie, China, 2019.
- [14] A. N. Ahmed, F. B. Othman, H. A. Afan et al., "Machine learning methods for better water quality prediction," *Journal of Hydrology*, vol. 578, no. 124084, 2019.
- [15] F. Othman, M. Alaaeldin, M. Seyam et al., "Efficient river water quality index prediction considering minimal number of inputs variables," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, pp. 751–763, 2020.
- [16] M. Rezaie-Balf, N. F. Attar, A. Mohammadzadeh et al., "Physicochemical parameters data assimilation for efficient improvement of water quality index prediction: comparative assessment of a noise suppression hybridization approach," *Journal of Cleaner Production*, vol. 271, article 122576, 2020.
- [17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [18] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, "Improving prediction of water quality indices using novel hybrid machine-learning algorithms," *Science of the Total Environment*, vol. 721, article 137612, 2020.
- [19] J. Y. Ho, H. A. Afan, A. H. El-Shafie et al., "Towards a time and cost effective approach to water quality index class prediction," *Journal of Hydrology*, vol. 575, pp. 148–165, 2019.
- [20] W. Shan, S. Cai, and C. Liu, "A new comprehensive evaluation method for water quality: improved fuzzy support vector machine," *Water*, vol. 10, no. 10, p. 1303, 2018.
- [21] M. Gaafar, S. H. Mahmoud, T. Y. Gan, and E. G. Davies, "A practical GIS-based hazard assessment framework for water quality in stormwater systems," *Journal of Cleaner Production*, vol. 245, article 118855, 2020.
- [22] T. H. Aldhyani, M. Al-Yaari, H. Alkahtani, and M. Maashi, "Water quality prediction using artificial intelligence algorithms," *Applied Bionics and Biomechanics*, vol. 2020, Article ID 6659314, 12 pages, 2020.
- [23] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2001.
- [24] A. Liaw and M. Wiener, "Classification and regression by random Forest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.
- [25] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 4-2, pp. 1486–1493, 2018.
- [26] R. K. Sahu, J. Müller, J. Park et al., "Impact of input feature selection on groundwater level prediction from a multi-layer perceptron neural network," *Frontiers in Water*, vol. 2, p. 46, 2020.
- [27] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22071–22080, 2019.
- [28] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: a new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.
- [29] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, vol. 454, Springer Science & Business Media, 2012.
- [30] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, and H. Liu, "Advancing feature selection research," *ASU feature selection repository*, vol. 1, no. 28, pp. 1–28, 2010.
- [31] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, and M. P. Mendes, "Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods," *Science of the Total Environment*, vol. 624, pp. 661–672, 2018.
- [32] T. Rajaei, S. Khani, and M. Ravansalar, "Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: a review," *Chemometrics and Intelligent Laboratory Systems*, vol. 200, article 103978, 2020.
- [33] 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.
- [34] L. Chamakura and G. Saha, "An instance voting approach to feature selection," *Information Sciences*, vol. 504, pp. 449–469, 2019.
- [35] S. G. Devi, "Analysing ground water quality in the regions of Kadapa District using supervised learning methods," in *International Conference On Computational And Bio Engineering*, Cham, 2019.
- [36] G. Jaihind, R. Ezhilarasie, and A. Umamakeswari, "Water quality monitoring and prediction of water quality at college premises using internet of things," *International Journal of Engineering and Advanced Technology*, vol. 3, pp. 53–57, 2019.
- [37] X. Li, J. Sha, and Z.-L. Wang, "Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake," *Environmental Science and Pollution Research*, vol. 25, no. 20, pp. 19488–19498, 2018.
- [38] O. Oyeboade, "Evolutionary modelling of municipal water demand with multiple feature selection techniques," *Journal of Water Supply: Research and Technology-AQUA*, vol. 68, no. 4, pp. 264–281, 2019.
- [39] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–18, 2017.
- [40] H. Bindra, R. Jain, G. Singh, and B. Garg, "Application of classification techniques for prediction of water quality of 17 selected Indian Rivers," in *Data Management, Analytics and Innovation*, pp. 237–247, Springer, 2019.
- [41] S. S. Shreem, S. Abdullah, and M. Z. A. Nazri, "Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm," *International Journal of Systems Science*, vol. 47, no. 6, pp. 1312–1329, 2016.
- [42] M. Mahdavi, M. Fesanghary, and E. Damangir, "An improved harmony search algorithm for solving optimization problems," *Applied Mathematics and Computation*, vol. 188, no. 2, pp. 1567–1579, 2007.

- [43] N. M. Nawi, A. S. Hussein, N. A. Samsudin, N. A. Hamid, M. A. M. Yunus, and M. F. Ab Aziz, "The effect of pre-processing techniques and optimal parameters selection on back propagation neural networks," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 3, pp. 770–777, 2017.
- [44] N. S. Sani, M. A. Rahman, A. A. Bakar, S. Sahran, and H. M. Sarim, "Machine learning approach for bottom 40 percent households (B40) poverty classification," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 8, no. 4-2, p. 1698, 2018.
- [45] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Cambridge, 2017.
- [46] Y. Luo, X. Cai, Y. Zhang, J. Xu, and X. Yuan, "Multivariate time series imputation with generative adversarial networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1603–1614, Montréal Canada, 2018.
- [47] C.-F. Tsai, M.-L. Li, and W.-C. Lin, "A class center based approach for missing value imputation," *Knowledge-Based Systems*, vol. 151, pp. 124–135, 2018.
- [48] B. Negara, R. Kurniawan, M. Nazri, S. Abdullah, R. Saputra, and A. Ismanto, "Riau forest fire prediction using supervised machine learning," *Journal of Physics: Conference Series*, vol. 1566, no. 1, article 012002, 2020.
- [49] S. Patro and K. K. Sahu, "Normalization: a preprocessing stage," 2015, <https://arxiv.org/abs/1503.06462>.
- [50] R. Bala and D. Kumar, "Classification using ANN: a review," *International Journal of Computational Intelligence Research*, vol. 13, no. 7, pp. 1811–1820, 2017.
- [51] M. M. Saritas and A. Yasar, "Performance analysis of ANN and Naive Bayes classification algorithm for data classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 7, no. 2, pp. 88–91, 2019.
- [52] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: performance calculation of dementia prediction by support vector machines (SVM)," *Informatics in Medicine Unlocked*, vol. 16, article 100200, 2019.
- [53] R. Gholami and N. Fakhari, "Support vector machine: principles, parameters, and applications," in *Handbook of Neural Computation*, pp. 515–535, Elsevier, 2017.
- [54] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [55] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability," 2012, <https://arxiv.org/abs/1206.1121>.
- [56] H. G. Debarba, M. E. de Oliveira, A. Lädermann, S. Chagué, and C. Charbonnier, "Augmented reality visualization of joint movements for physical examination and rehabilitation," in *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, Tuebingen/Reutlingen, Germany, 2018.
- [57] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.
- [58] R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, p. 81, 1986.
- [59] H. Zhou, J. Zhang, Y. Zhou, X. Guo, and Y. Ma, "A feature selection algorithm of decision tree based on feature weight," *Expert Systems with Applications*, vol. 164, article 113842, 2021.
- [60] S. M. Lundberg, G. Erion, H. Chen et al., "From local explanations to global understanding with explainable AI for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [61] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
- [62] F. Wang, Y. Wang, K. Zhang, M. Hu, Q. Weng, and H. Zhang, "Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation," *Environmental Research*, vol. 202, article 111660, 2021.
- [63] M. Hameed, S. S. Sharqi, Z. M. Yaseen, H. A. Afan, A. Hussain, and A. Elshafie, "Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia," *Neural Computing and Applications*, vol. 28, no. S1, pp. 893–905, 2017.
- [64] G. Hayder, M. I. Solihin, and H. M. Mustafa, "Modelling of river flow using particle swarm optimized cascade-forward neural networks: a case study of Kelantan river in Malaysia," *Applied Sciences*, vol. 10, no. 23, p. 8670, 2020.
- [65] R. M. Abdul, N. S. Sani, R. Hamdan, O. Z. Ali, and B. A. Abu, "A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group," *PLoS One*, vol. 16, no. 8, article e0255312, 2021.