EGE AKADEMİK BAKIŞ / *EGE ACADEMIC REVIEW*

Cilt 23 • Sayı 2 • Nisan 2023
SS. 265/278
Doi: 10.21121/eab.1252167
Başvuru Tarihi: 17.02.2023 • Kabul Tarihi: 01.03.2023

Article Type: *Research Article*

Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms for Better Sustainability

Mustafa YURTSEVER[1], Murat EMEÇ[2]

ABSTRACT

Water is one of the most important resources for human life and health. Global climate change, industrialization and urbanization pose serious dangers to existing water resources. Water quality has traditionally been predicted by expensive, time-consuming laboratory and statistical analysis. However, machine learning algorithms can be applied to determine the water quality index in real time efficiently and quickly. With this motivation, a dataset obtained from the Kaggle website was used to classify water quality in this research. Some features were found to be empty in the data set. Traditional methods (drop, mean imputation) and regression method were applied for null values. After the null values were completed, RF, Adaboost and XGBoost were applied for binary classification. Gridsearch and Randomsearch methods have been applied in hyper parameter optimization. Among all the algorithms used, the SXH hybrid method created with the Support Vector Regression (SVR) and XGBoost methods showed the best classification performance with 99.4% accuracy and F1-score. Comparison of our results with previous similar studies showed that our SVR XGboost Hybrid (SXH) model had the best performance ratio (Accuracy, F1-score). The performance of our proposed model is proof that hybrid machine learning methods can provide an innovative perspective on potable water quality.

Keywords: Water Quality Index, Machine Learning, Classification, Imputation Methods, Regression.

JEL Classification Codes: C38, Q25, Q56

Referencing Style: APA 7

INTRODUCTION

Water is the primary resource for all human, animal, and plant life. Although its primary use was potable water, it was also used for industry, agriculture, and commerce. One of the essential elements for maintaining human life is water. Water is necessary for the continuation of life events in our bodies. It is also the primary energy source and provides life force by producing electrical and magnetic energy in every body cell. Some elements and compounds in potable water are necessary for the human body, provided they are not more than necessary. Because they are the elements that make up the structure of cells, the basic unit of living things. The biological solvent effect of water ensures that vitamins and minerals are transported and dissolved in the body. It also plays a role in regulating body temperature, functioning of the kidneys, and cleansing the body. Access to potable water is essential to prevent various water-borne diseases. The fact that the earth is covered with 71% water gives the appearance that there is plenty of water in the world, but the rate of potable water is very low. Only 1.2% can be used as potable water (National Geographic, 2022).

Unsafe potable water causes more than 1.5 million deaths from diarrhea each year, mostly infants and young children (WHO, UNICEF, World Bank, 2022). Due to its vital, economic, and strategic importance, water continues its potential to be the most discussed topic of today and the near future. Globalization, global climate change, industrialization, and large amounts of domestic waste seriously threaten existing water resources. Another issue in the water use of the sectors is the pollution of the existing water potential. Fresh water, which has already become limited and valuable, is irreversibly polluted due to not treating industrial and domestic wastes. Obtaining clean water due to purification from contaminated water requires great costs. Potable water refers to water suitable for human consumption. It should be in a structure that will not endanger human health, and that is not contrary to the working principles of metabolism. After being collected from rivers, lakes, and wells, potable water is presented to the consumer through various processes.

According to the WHO (2022), water producers are responsible for ensuring the safety and quality of their products. The US Environmental Protection Agency

[1] Dokuz Eylul University, IT Department, Buca, 35390, Izmir, TÜRKİYE, mustafa.yurtsever@deu.edu.tr
[2] Istanbul University, IT Department, Fatih, 34116, Istanbul, TÜRKİYE, murat.emec@istanbul.edu.tr

(EPA) establishes rules and guidelines for a wide range of contaminants, such as chemicals and bacteria that can cause disease, that is present in public drinking water supplies. It sets legal limits for more than 90 pollutants in potable water. Potable water quality rules guarantee that consumers can access reliable, sufficient, and secure drinkable water. With industrialization and urbanization, it is one of the primary responsibilities of city managers to deliver healthy potable water to citizens. Water quality refers to the standards for water's chemical, physical and biological properties (Liou, 2004). The abundance and complexity of the variables that define water quality make it difficult to measure and calculate water quality. Therefore, water quality indexes have been created to assess the acceptability of water for various uses. This idea compares the appropriate standards and the water quality measure. Non-water professionals can easily understand the results of the water quality index (WQI), which presents a significant amount of water quality data in a single figure (Abed et al., 2022). The water quality index combines different parameter values in different units and sizes into a single dimensionless number with the help of different aggregation functions, sub-indexes, and weighting factors.

Traditional testing has numerous flaws, but we must overcome them if we are to ensure that human water is safe and aquatic ecosystems are stable. The laboratory test can measure multiple parameters and give accurate results, but the processes are very long. Artificial intelligence (AI) approaches are becoming more and more common for accurately and quickly detecting and monitoring water quality in real time. Unlike traditional models, AI models offer better solutions for nonlinear problems. AI has many uses in different fields. AI technology is frequently used in many business areas, such as medicine, health, education, military, agriculture, economy, finance, automotive, telecommunications, mining, media, banking, and many more. However, due to their unique characteristics, it is difficult to research the quality of different water types (seawater, groundwater, fresh water, etc.). Machine learning (ML) methods, a sub-branch of AI, are seen as an effective tool to overcome these difficulties (Zhu et al., 2022).

One of the problems that have a significant impact on performance when using ML methods is missing data. Before using the data for ML models, making decisions about missing values is necessary. Missing data refers to the lack of observations in the data set, which is expected but cannot be recorded. Missing values in real-world data can occur for a number of reasons, such as unsaved observations and corrupted data. Work on missing data must be done in advance because ML algorithms do not accept data with missing values. Also, completely ignoring missing values can lead to biased results. Ignoring missing values can become a minimum-size dataset where ML applications will be meaningless. Missing data is considered insignificant when it makes up less than 1% of the total data. Up to 5% of a rate is regarded as manageable. However, rates above the 5% cutoff and close to 15% call for diverse treatment strategies. Filling in the missing value is one of the primary methods. Mean value imputation or regression imputation can be used to fill in the missing value. There are also modern techniques, such as deep learning and expectation approaches. Studies also use hybrid methods (Zhang et al., 2020; Rani et al., 2021).

The rest of this article is divided into the following sections: First, a literature review was conducted, and related studies were presented. Then, the material and method are described. Next, the recommended hybrid model is presented in Chapter Proposed Hybrid Model. The final section presents results, comparisons, and discussion.

RELATED WORK

Most studies in the literature use traditional laboratory analyses and data analysis to measure the quality of water. Some recent studies have determined water potability with ML methods. AI, ML, and deep learning methods are used on very different data sets in various fields. A machine may mimic human behavior thanks to AI technology. A subfield of AI called ML enables computers to learn from past data without explicit programming automatically. Deep learning is an AI technique that trains computers to analyze data in a way similar to how the human brain does it.

Consumers' health can be adversely affected by the quality of potable water. Potable water quality is mainly affected by the quality of the extracted water and its processing, distribution, and preservation processes before it reaches the consumer. Therefore, effective and rapid potable water quality assessment approaches gain importance when economic developments, technological developments, and the health of the increasing population are considered. Using AI and ML algorithms for potable water quality prediction can lead to a more sustainable approach to managing water resources. By providing real-time monitoring, early detection, and optimized treatment options, these algorithms can help ensure the availability of safe and clean water for communities worldwide.

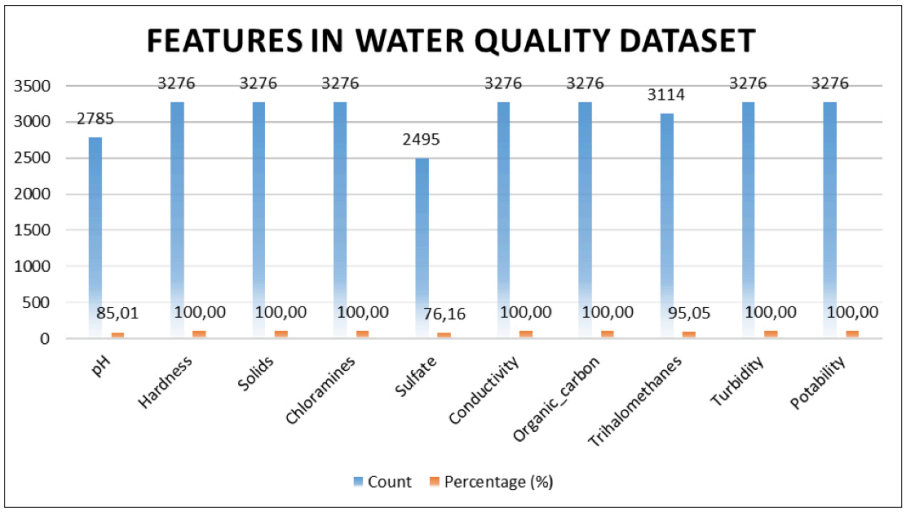

Figure 1. Distribution of features in the water quality dataset

Chafloque et al. (2021) used a neural network-based algorithm to attempt to predict if water is fit for human use. The model they used achieved a 70% accuracy rate. Their study ignored null values by removing them from the data set. In another study using different ML methods to estimate water drinkability, the k-nearest neighbor and support vector machines LASSO LARS and stochastic gradient descent gave the best results according to different evaluation parameters. The study excluded null values from the data set (Kaddoura, 2022). In another study using the same data set, ML methods J48, Naive Bayes, and multi-layer percepton (MLP) were used to predict water quality. They filled null values with mean and median. MLP gave the highest accuracy (Abuzir and Abuzir, 2022).

Xin and Mou (2022) stated that sulfate, pH, solids, and hardness are the most critical factors in predicting water quality. The ML methods that give the best results in their work are XGBoost, CatBoost, and LGBM, respectively. Fen et al. (2021) obtained 75.83% overall prediction accuracy with the decision trees algorithm. Using ML methods, Patel et al. (2022) achieved 81% accuracy with Random Forest and Gradient Boost. They used the mean method for null values.

Azrour et al. (2022) developed a model that can predict the water quality index and, subsequently, the water quality class. The neural networks predicted the water quality class with an accuracy exceeding 85%. Dilmi and Ladjal (2022), using LSTM, one of the deep learning methods, used different feature extraction techniques to improve classification quality. They found an accuracy rate of 99.72% in the study.

Ahmed et al. (2019) tested different ML algorithms using four parameters: temperature, turbidity, pH, and total dissolved solids. MLP stood out as the best method, with 85% accuracy. Aldhyani et al. (2020) state seven important parameters, such as conductivity, pH, and nitrate, in the data set containing 1679 samples. Support Vector Machines gave better results than KNN and Naive Bayes algorithm with the highest accuracy rate of 97%.

MATERIALS AND METHODS

This section describes the materials (dataset), hardware, and software tools used in our proposed hybrid ML model and the methods applied to fill in the missing value and the binary classification method.

Description of Dataset Used in Our Work

The search for a suitable dataset for this research started with reading similar studies from the past, browsing the internet, and reviewing many other scientific sources. In the first investigation, we collected several datasets containing water quality parameters and started to analyze them. Finally, we selected the dataset "Water Quality" from the Kaggle website for water quality estimation, which was then used to train and test the model (Kadiwal, 2022). We chose the dataset to include important parameters used in water quality measurement. The dataset contains a total of 3276 data.

Figure 1 presents the distribution of the features in the dataset. Of the ten key features presented in Figure 1, 9 are for water properties, and one is for the "potability" of water. The "potability" value is either 0 or 1. 0 represents non-potable water, and 1 represents potable water.

The pH value is a unit of measurement expressing the acidity or alkalinity of any water. Water hardness is usually the amount of calcium and magnesium dissolved in the water. Total Dissolved Solids (TDS) include minerals, cations, anions, heavy metal ions, and small amounts of organic matter dissolved in water and cannot be retained by simple filtration methods such as sand filters. The higher the TDS in the water, the more foreign matter there is. Chloramines are formed by adding ammonia to chlorine during drinking water treatment. Sulphate is one of the most important lines that rain dissolves. The high isolation powers in our water can be detrimental when combined with shield and army, the two most common hardness components. Conductivity is a property that determines the purity of water. The lower the conductivity of the water, the fewer the ions in it. Organic carbon in spring water comes from natural and artificial sources, including decaying organic matter. Drinking water is safe for ingestion by humans. Its value is 0 or 1.

WQI is measured according to various parameters of water. Water quality index calculation methodology is listed as follows (Brown et al., 1972).

1. Collecting data on various physicochemical water quality parameters

2. Calculation of the proportionality constant "K" ("si" is the standard value of the nth parameter)

$$K = (1/(1/\sum_{n=1}^{n}(s_n)) \quad (1)$$

3. Calculate the quality grade for the nth parameter (qn) with n.

$$q_n=100\ \{(v_n-v_{io})/(s_n-v_{io})\} \quad (2)$$

4. Calculation of weight of units for parameters

$$W_n=(k/s_n) \quad (3)$$

5. Water Quality Index Calculation Formula

$$WQI = \frac{\sum_{i=0}^{n} qi*wi}{\sum_{i=0}^{n} wi} \quad (4)$$

In the proposed model, water quality measurement is assessed using ten critical water quality indicators from the preferred dataset. Reference levels of the water quality index are classified by the World Health Organisation (WHO). According to (Brown et al., 1972), the index level must be less than 50 for water to be potable.

The figure 2 presents the data of 1,998 for non-potable waters and 1,278 for potable waters in the dataset. The figure shows that the data set is unbalanced.

The dataset contains a high proportion of features with missing values. Figure 3 shows the distribution of missing values for the three features. Missing values were approximately 15% for "pH," 24% for "Sulphate," and 5% for "Trihalomethanes" property.

Data Preprocessing

As part of data preprocessing, non-numeric data is converted to numeric numbers. In addition, the duplicate data is deleted, and only the necessary data is kept. In data preprocessing, NULL values are primarily detected. The algorithm must be able to function without any missing data because null values signify missing data. The method can also produce more accurate results by substituting null values. As observed in Figure 3's graphic, the values for "pH," "Sulphate," and "Trihalomethanes" are NULL.

After processing the missing data, non-numeric entries in the data set were converted to numeric values. The next stage of data preparation is data normalization. The normalization method, a standard scaler, was used to place the data in the range [0, 1] (Kaushik et al., 2019; Graf et al., 2022). After the normalization step, the (0,1) data are transformed into a TensorFlow, labeled for features and classification. Following the TensorFlow step, the input and output parameters of the learning model are defined. Finally, the data set is divided into two subsets as training and testing.

Briefly, our data preprocessing consists of the following stages:

- Storing and checking the data set in computer memory
- Detection and processing of missing data
- Conversion of nominal data to numerical data
- Normalize data using the standard scaler
- Subdivision into subsets for the Water Quality dataset:
    - o Training set: 2,620
    - o Test set. 656

The training and test data sets are separated by 80% and 20%, respectively.

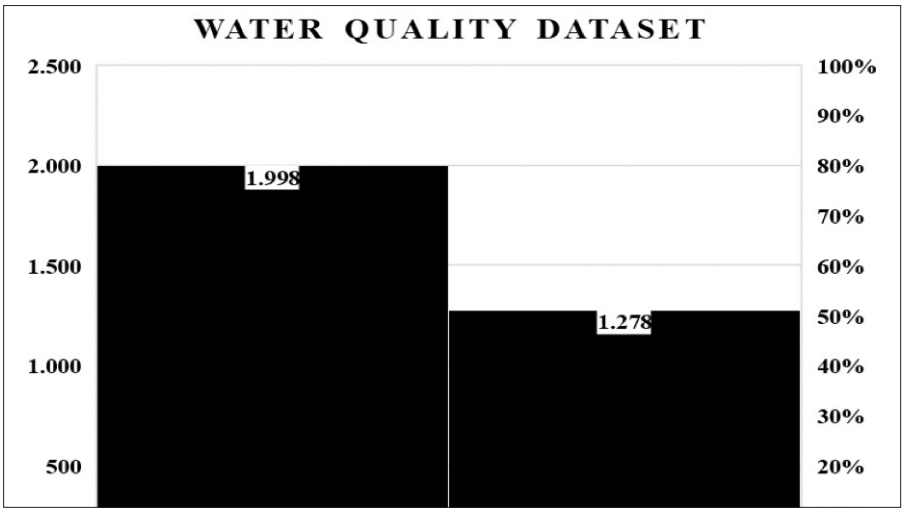

Figure 2. Classification of the potability of the water in the data set used

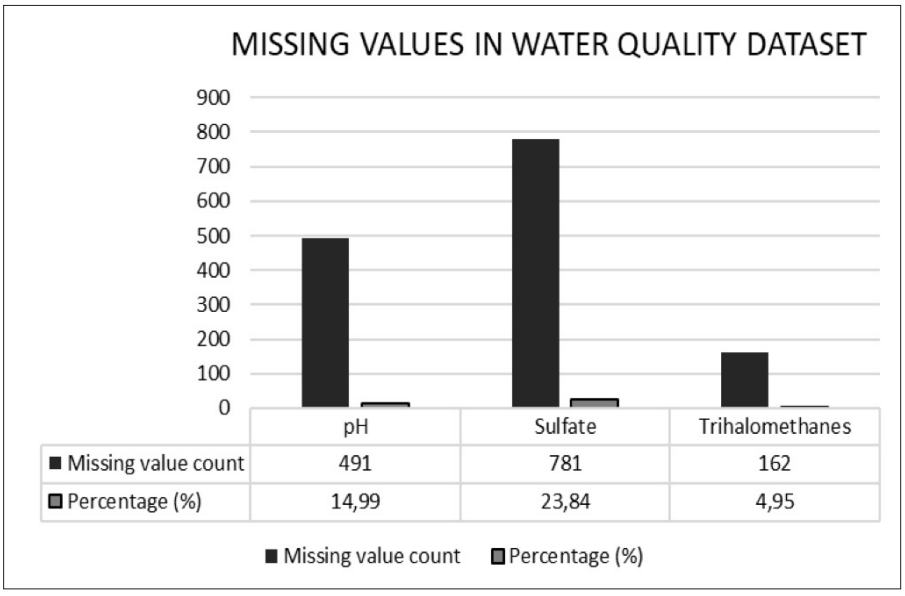

Figure 3. Distribution of missing values in the data set used

Feature Engineering and Importance

Feature engineering is applied to handle missing data and improve model prediction accuracy. The correlation matrix method was used to identify the connections between the features.

Table 1 presents the correlations representing the relationships between the features in the dataset. As a first step in feature engineering, the missing values of the three features in Figure 3 were determined. Then, a correlation matrix was created for each feature with missing values in these features. Finally, the importance scores of the features in the dataset were obtained.

Table 2 shows the significance scores of the features of the Random forest and XGBoost algorithms in the data set. When the table is examined in detail, the importance degrees of the "Sulfan" and "pH" features are in the first two places in both methods.

Prediction of Missing Values Using Support Vector Regression (SVR)

Water quality is sensitive data; it may be insufficient to impute water data using other methods of imputing missing values (e.g., mode, mean, median). Therefore, there is a need for an innovative method for filling in missing values. In the first step of our proposed method, missing features were detected. These features are; "pH," "Sulfate," and "Trihalomethanes." Next, the features with positive correlation were determined in the correlation matrix of the features in Table 2.

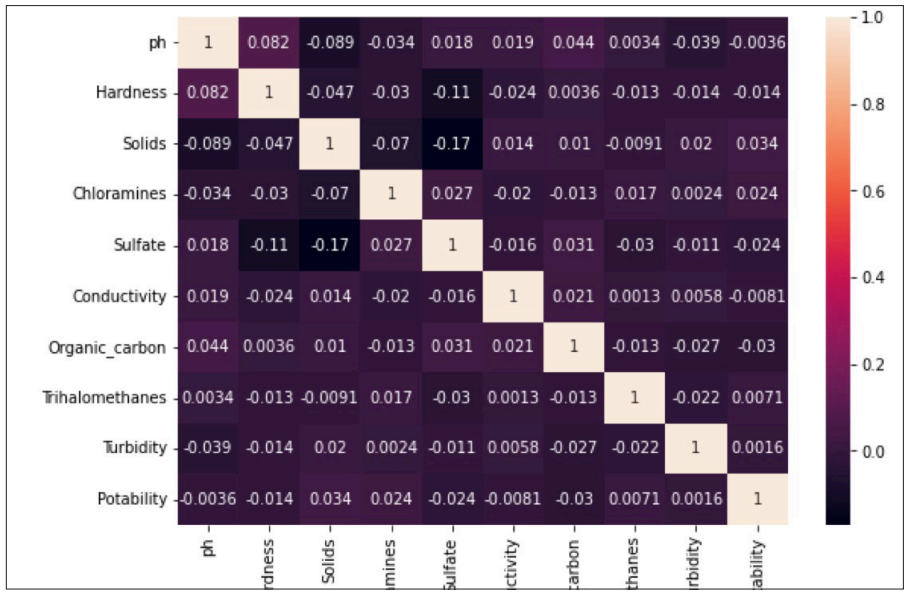

Table 1. Correlation matrix of Water Quality attributes

Table 2. Significance scores of water quality dataset attributes

Feature(s)	Random Forest Feature importance (%)	XGBoost Feature importance (%)
Sulfate	26.8	13.8
pH	20.9	13.6
Chloramines	13.4	13.1
Hardness	10.5	12.5
Solids (TDS)	8.9	12.7
Trihalomethanes	5.6	8.1
Conductivity	4.9	9.1
Turbidity	4.6	8.3
Organic Carbon	4.5	8.9

Positively correlated features;

- For "pH": ['Hardness,''sulfate,''Conductivity,' 'Organic_carbon,''Trihalomethanes]
- For "Sulphate": ['pH, 'Organic_carbon,' 'Chloramines']
- For "Trihalomethanes": ['Chloramines,' 'Conductivity']

Figure 4 shows the step-by-step process of filling in the missing values. The steps in the sequence were repeated until all columns were filled.

Binary Classification

In this stage, ML methods were trained for binary classification to distinguish between potable and non-potable water conditions in the water quality dataset.

The proposed classification method architecture for measuring potable water quality is shown in Figure 5. The classification architecture in the figure is designed by using hyperparameters (n_estimators, max_depth, etc.) of the learning model to obtain the best results. Hyperparameters have been adjusted for best results.

A supervised learning method was adopted in the research. The water quality dataset labels potable water as potable (1) or non-potable (0). Our learning model is first trained with the training set, then binary classification prediction is performed with the test data.

PROPOSED HYBRID MODEL (SVR+XGBOOST)

ML algorithms have been proposed for the classification model of potable water quality. In this direction, the main objective of the study is to predict the intended labelled data with the best performance by training the data in

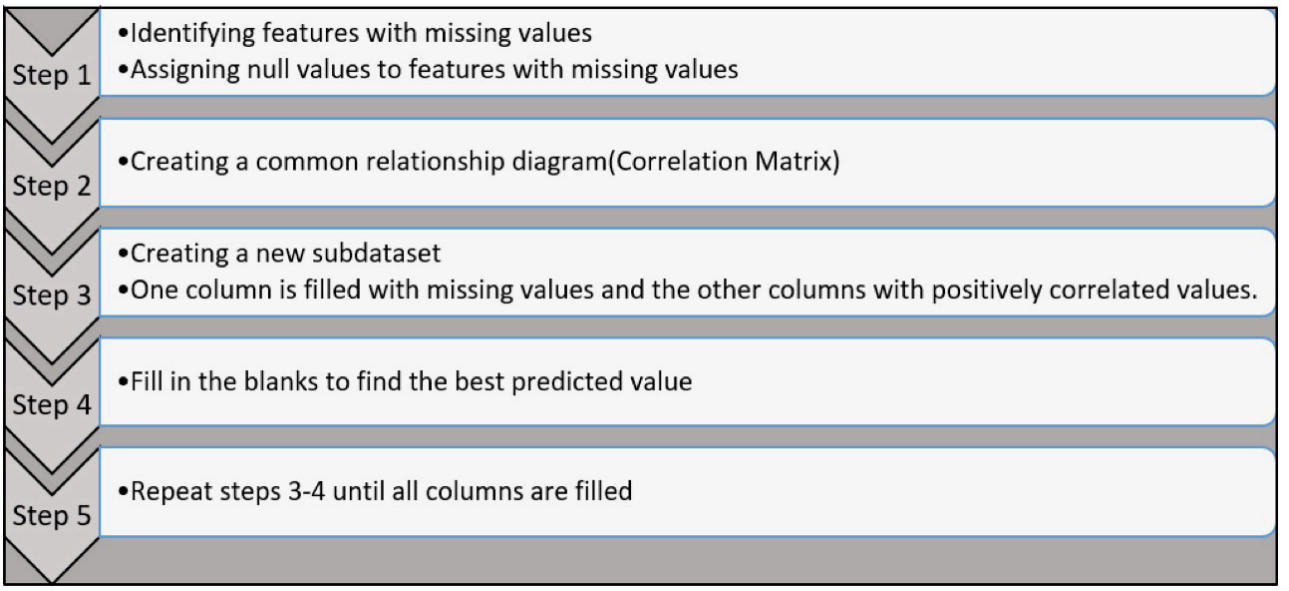

Figure 4. The steps followed in the proposed approach to fill in the missing value

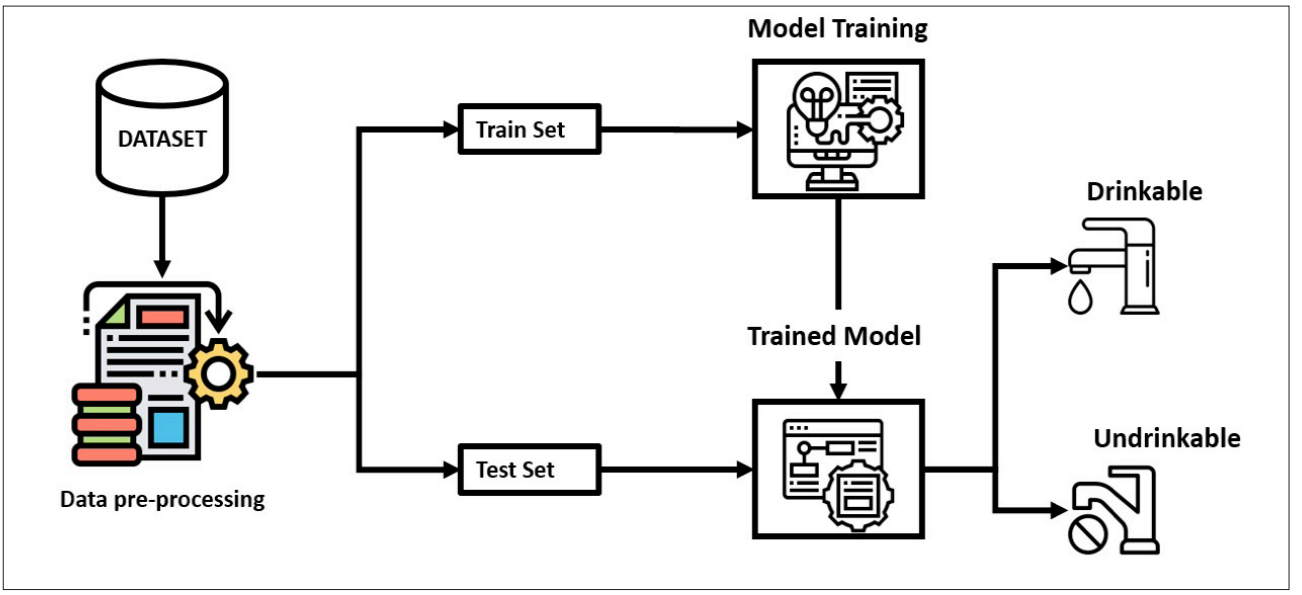

Figure 5. Classification architecture for potable water

the training set to recognise input-output mappings. Inferential functions are then produced after that. These functions then translate the new unlabeled data into the correct class in a subsequent stage (testing process) (Kaushik et al., 2019; Graf et al., 2022).

Vapnik presented the current fundamental SVM method in the early 1990s (Vapnik, 1998). Support vector machines (SVMs) assess classification and regression analysis data. SVMs are algorithms developed for supervised learning. It contributes to the learning model by analyzing large amounts of data to find relationships between data and detecting relationship states. Finding the ideal hyperplane with the smallest distance to all data points is the goal of SVM regression. As previously noted, SVM can be used for various issues, including classification, clustering, and regression issues. For the estimation of water quality and parameters, successful results have been obtained in previous studies using the support vector regression algorithm (Wang et al., 2020; Wang et al., 2011). Within the scope of our study, support vector regression method was applied to fill the missing parameters in the water quality data set. The Grid Search method is adopted as the hyperparameter for support vector regression parameters. Grid search parameters are set as follows.

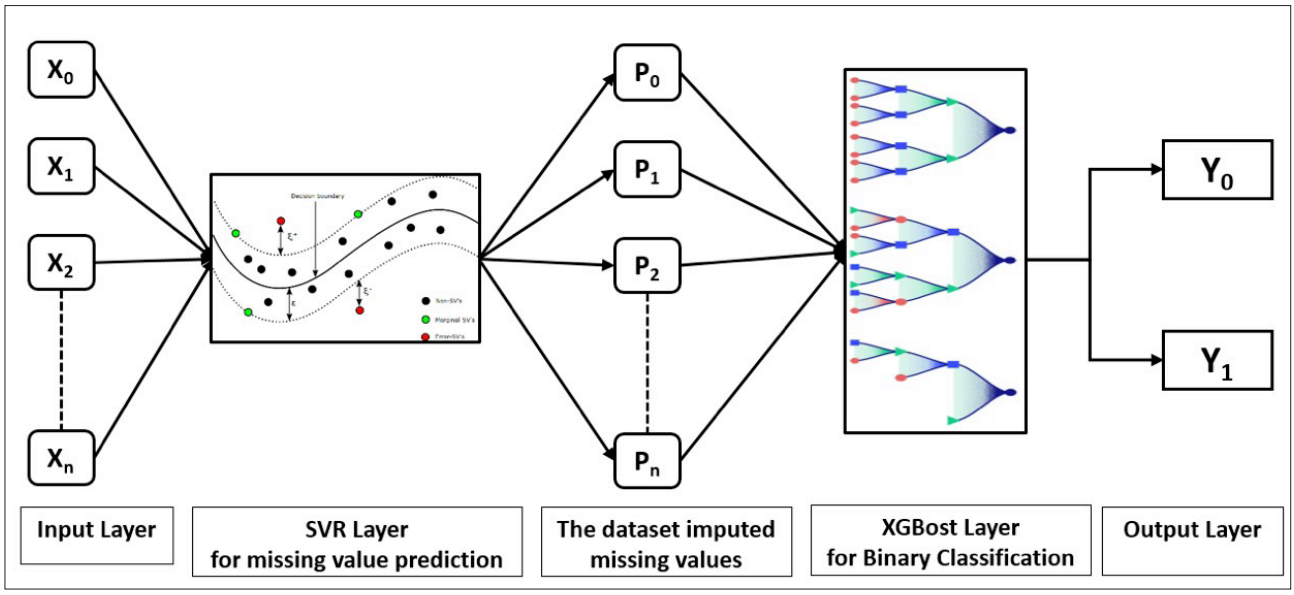

Figure 6. Proposed model architecture

• kernel('sigmoid','linear','poly')

• degree (1,7,9,2)

• gamma ('scale',' auto')

• C (1.2,1.3,1.4,1.7

The best parameters for the SVR model:

- kernel: 'poly'
- degree: 2
- gamma: 'auto'
- C: 1.2

Gradient Boost Machines (GBM), one of the most effective algorithms for supervised learning, is implemented in XGBoost (XGB), one of its variants. Additionally, it can be used to address issues regarding regression and classification. The XGB algorithm does well in many ML challenges. For various predictive modeling use cases, logistic regression modeling appeared to be the best approach. However, as time went on, it lost out in the literature to XGBoost. Despite its strong algorithm adaptability, it learns thanks to parallel and distributed computing quickly and provides expert memory use. In order to make the final prediction choice, XGBoost uses a number of different models to provide an output, making it an ensemble learning technique. In the decision process of the classification and training model architecture for the detection of potable water quality, a detailed methodological survey of studies in the literature was performed. Our potable water hybrid model was chosen after comparisons with single algorithm models (Kaddoura, 2022; Chafloque et al., 2021) compared to previous hybrid models based on superior performance results (Zhang et al., 2020; Rani et al., 2021). SXH consists of 5 layers, as shown in Figure 6: Input, SVR, Missing Value, XGBoost, and Output layers. SVR algorithm is used for the estimation of missing values. XGBoost algorithm was used for classification.

Briefly, the stages of our proposed SXH model are as follows:

1. Ten inputs ($X_{0,}$ ... X_n) in the water quality dataset;
2. Estimation of missing values in the SVR layer (P_0);
3. Preparation of the new dataset with the predicted values for the input of the XGBoost layer;
4. Classification in the XGBoost layer;
5. Modeling of decision transmission (potability, non-potable) to the output layer ($Y_{0'}$ ... Y_n).

Hyperparameter tuning of the proposed model has been accomplished through random search and grid search, the programmer's heuristic, and previous experiments and literature reports (Xie et al., 2019).

The random search and grid search parameters of the proposed model are listed below:

• n_estimators: (400 ,800,1000)

• learning_rate (0.5,1)

• max_depth': range (1, 11, 2)

Table 3. Random forest confusion matrix for Grid search hyperparameter tuning

Missing value imputation method (s)		Predicted (1)	Predicted (0)
DROP	Actually (1)	197	34
	Actually (0)	100	72
MEAN	Actually (1)	382	30
	Actually (0)	97	174
REGRESSION	Actually (1)	387	25
	Actually (0)	39	205

Table 4. AdaBoost confusion matrix for Grid search hyperparameter tuning

Missing value imputation method (s)		Predicted (1)	Predicted (0)
DROP	Actually (1)	190	41
	Actually (0)	126	46
MEAN	Actually (1)	396	16
	Actually (0)	151	93
REGRESSION	Actually (1)	404	8
	Actually (0)	23	221

The best parameters for the model:

Random search for: (400, 1, 1)

For Grid search: (1000, 0.5, 1)

RESULTS AND COMPARISON OF PREVIOUS WORK

This section will first define the performance measures for the suggested models. Second, filling in missing values and the results of our binary classification models will be presented. Then the performances of the models will be compared. Finally, our results will be compared with previous water quality studies using different methods. As in previous studies, four popular measures were used to evaluate performances. These are accuracy (1), precision (2), recall (3), and f1 score (4). In addition, the Confusion Matrix table, which presents the summary of the estimation results in a classification problem, is used.

RESULTS

After applying the dataset to each of the Random Forest, AdaBoost, and SXH algorithms, confusion matrices (Tables 3-5) were generated.

When Table 3 is examined in detail, the complexity matrix is seen according to the imputation methods of 3 different missing values. According to Table 3, the management obtained the highest REGRESSION missed value imputation using the RF model, the potable water samples (1), and predicted potable water samples with 387. On the other hand, the closest value to REGRESSION with 382 was obtained by the MEAN missing value method.

When Table 4 is examined, the imputation methods assign the missing values, and the predicted potable water samples using the AdaBoost model were obtained by the REGRESSION imputation method as 404. The value close to the highest value was obtained from the MEAN missing value imputation method with 396.

Table 5 presents the results of the proposed SXH method for estimating potable water quality. According

Table 5. SXH confusion matrix for Grid search hyperparameter tuning

Missing value imputation method (s)		Predicted (1)	Predicted (0)
DROP	Actually (1)	162	69
	Actually (0)	84	88
MEAN	Actually (1)	375	37
	Actually (0)	97	147
REGRESSION	Actually (1)	412	0
	Actually (0)	3	241

Table 6. Random forest results for binary classification

Missing value imputation method (s)	Category	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DROP	Model	66,74938	67,01052	66,74938	64,88053
	Potable	79,82196	66,32997	85,28139	74,62121
	Non-Potable	57,35608	67,92453	41,86047	51,79856
MEAN	Model	81,40556	81,94947	81,40556	80,79328
	Potable	90,25974	79,74948	92,71845	85,74635
	Non-Potable	74,13333	85,29412	64,20664	73,26316
REGRESSION	Model	**90,24390**	**90,20731**	**90,24390**	**90,18132**
	Potable	92,21184	90,84507	93,93204	92,36277
	Non-Potable	88,35821	89,13043	84,01639	86,49789

to Table 5, potable water samples (1) and predicted with the proposed SXH model obtained the highest predictive value with a score of 412. All potable water samples were predicted to be true with the proposed model. However, non-potable(0) 3 water samples were predicted as potable. Results from the tests are shown in Table 6, Table 7, and Table 8, respectively. The tables present the classification metrics obtained from each ML method.

When Table 6 is examined in detail, it is seen that the highest performance for the RF model is obtained with the REGRESSION method, one of the null value filling methods. The accuracy and F1-score results obtained with the REGRESSION method were 90.24% and 90.18%, respectively. With the MEAN method closest to this performance, accuracy and F1-Score scores of 81.40% and 80.79% were obtained, respectively.

When Table 7 is examined, the highest performance for the Adaboost model was obtained with the REGRESSION method, one of the null-filling methods. The accuracy and F1-score obtained by the REGRESSION method were 95.27% and 95.24%, respectively. On the other hand, the accuracy results of MEAN and DROP imputation methods showed poor performance at 74.54% and 58.56%, respectively.

When Table 8 is examined in detail, the highest performance for the SXH model was obtained with the REGRESSION method, one of the null-filling methods. The accuracy obtained with the REGRESSION method and the F1-score score of 99.54% was obtained. RF and AdaBoost results are presented in Table 6 and Table 7. The proposed SXH results in Table 8 outperformed the results of the RF models in Table 6 and the AdaBoost models in Table 7. These performance results are the main reasons we recommend the SXH method.

Random search and Grid search hyperparameter results are presented in Table 9. When the table is examined in

Table 7. Adaboost results for binary classification

Missing value imputation method (s)	Category	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DROP	Model	58,56079	57,03100	58,56079	54,98061
	Potable	74,21384	60,12658	82,25108	69,46984
	Non-Potable	48,36066	52,87356	26,74419	35,52124
MEAN	Model	74,54268	77,20280	74,54268	71,46662
	Potable	93,85797	72,39488	96,11650	82,58603
	Non-Potable	61,82048	85,32110	38,11475	52,69122
REGRESSION	Model	**95,27439**	**95,31767**	**95,27439**	**95,24170**
	Potable	97,50390	94,61358	98,05825	96,30513
	Non-Potable	93,14456	96,50655	90,57377	93,44609

Table 8. SXH results for binary classification

Missing value imputation method (s)	Category	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
DROP	Model	62,03474	61,66987	62,03474	61,76621
	Potable	64,43299	65,85366	70,12987	67,92453
	Non-Potable	59,80861	56,05096	51,16279	53,49544
MEAN	Model	79,57317	79,61361	79,57317	78,83460
	Potable	87,58389	79,44915	91,01942	84,84163
	Non-Potable	72,90503	79,89130	60,24590	68,69159
REGRESSION	**Model**	**99,54268**	**99,54599**	**99,54268**	**99,54210**
	Potable	100,00000	99,27711	100,00000	99,63724
	Non-Potable	99,08953	100,00000	98,77049	99,38144

detail, the results obtained with the Grid search method showed better performance than those obtained with Random Search in all accuracy, Precision, Recall, and F1-score metrics. For this reason, the Grid Search method has been adopted in our proposed model.

Outcomes from Our SXH Model Compared to Earlier Works

This section presents a table comparing our SXH model with previous drinking water quality studies using ML methods, considering accuracy and F1-score parameters. According to the results obtained, it was determined that the use of SXH outperformed RF and Adaboost in this study.

In Table 10, comparisons of the results of our hybrid method with other studies using the "Water Quality" data set we used within the scope of the study are presented. The table is detailed by Fen et al. (2021); by deleting the missing values from the data set, they obtained 86.67% and 85.78% accuracy and F1-score, respectively, with the Extra trees classifier method. Filling in the missing values with the mean value, Patel et al. (2022) achieved an acceptable accuracy rate of approximately 80% with RF and F1-score. Other studies performed poorly, staying below 80% accuracy.

In the hybrid method we recommend, it is seen that the accuracy and F1-score score of 99.64% is approximately

Table 9. Comparison of results of hyperparameter methods

Method(s)	Model(s)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Random Search	Our Proposed -SXH	99,39024	99,39611	99,39024	99,38920
Grid Search	Our Proposed -SXH	99,54268	99,54599	99,54268	99,54210

Table 10. Comparison Table of Our SXH Model Results with Previous Works

Author(s)	**Proposed Model**	**Accuracy (%)**	**F1-Score (%)**
Kaddoura, 2022	ANN	-	63.90
Chafloque et al., 2021	Neural Network	69.00	-
Abuzir and Abuzir, 2022	Multi-layer perceptron (MLP)	-	75.90
Xin and Mou, 2022	XGBoost	81.51	80.78
Fen et al., 2021	Extratrees classifier	86.67	85.78
Patel et al., 2022	Random Forest	81.00	81.50
Our Proposed	SXH	**99.54**	**99.54**

14% better than the highest study. This success of our study was realized with the effect of the null value regression and binary classification method we used.

CONCLUSION

As a human right, safe and clean potable water and health protection are vital for fully enjoying the right to life. However, a number of pollutants reduce the quality of drinkable water. Water quality indices are approaches that simplify the expression of potable water status. In this way, it enables individuals and institutions not experts in water quality to obtain information about the water quality status and use this data comfortably, quickly, and efficiently. Predicting potable water quality is an essential aspect of ensuring its sustainability. AI and ML algorithms can be powerful tools for this task.

This study presents a new hybrid model that predicts the drinking water quality index. First, the performance of the methods that fill in the missing values in the data set and then the performance of the classification algorithms are measured. Finally, a hybrid model was created with a combination of algorithms that gave the best results.

The SXH model we used in the study performed best in all binary potable water classifications according to accuracy and F1-score values. Our SXH model was also compared with other potable water prediction models using various methods. By comparison, the accuracy performance of our proposed model was found to outperform the closest run by about 13%. It also performed about 14% better than the most in the other performance metric, the F1-score. The results show that the hybrid method is a very successful model for potable water estimation. It is aimed to apply the method we propose for further studies to a different data set that includes other parameters used in measuring water quality. In addition, water analysis organizations or companies can apply our proposed hybrid model to the data they have obtained to determine potable water quality.

In the administration of water resources, ecological restoration and establishment of mechanisms that will rearrange water consumption according to industry, agriculture, and drinking water needs are important. In this framework, existing and planned projects have to be reviewed.

REFERENCES

Abed, B. S., Farhan, A. R., Ismail, A. H., & Al Aani, S. (2022). Water quality index toward a reliable assessment for water supply uses: a novel approach. *International Journal of Environmental Science and Technology*, *19*(4), 2885-2898.

Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. *Water Quality Research Journal, 57(3),* 152-164.

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)*, 2210.

Aldhyani, T. H., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics*, 2020.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. Modeling *Earth Systems and Environment, 8(2),* 2793-2801.

Brown, R. M., McClelland, N. I. Deininger R. A. and O'Connor, M. F. (1972). Water Quality Index-Crashing, the Psychological Barrier, Proc. 6th Annual Conference, Advances in Water Pollution Research, pp 787-794.

Chafloque, R., Rodriguez, C., Pomachagua, Y., & Hilario, M. (2021, September). Predictive Neural Networks Model for Detection of Water Quality for Human Consumption. In 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 172-176). IEEE.

Dilmi, S., & Ladjal, M. (2021). A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques. *Chemometrics and Intelligent Laboratory Systems, 214*, 104329.

Fen, L., Lei, Z., & Ting, C. (2021, November). Study on Potability Water Quality Classification Based on Integrated Learning. In 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (pp. 134-137). IEEE.

Graf, R., Zeldovich, M., & Friedrich, S. (2022). Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study. *Biometrical Journal.*

Kaddoura, S. (2022). Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. *Sustainability, 14(18),* 11478.

Kadiwal, A. (2022) Water Quality [Dataset]. https://www.kaggle.com/adityakadiwal/water-potability. Accessed on 24 December 2022

Kaushik, P., Gupta, A., Roy, P. P., & Dogra, D. P. (2019). EEG-based age and gender prediction using deep BLSTM-LSTM network model. *IEEE Sensors Journal, 19(7),* 2634-2641.

Liou, S. M., Lo, S. L., & Wang, S. H. (2004). A generalized water quality index for Taiwan. *Environmental monitoring and assessment, 96*, 35-52.

Patel, J., Amipara, C., Ahanger, T. A., Ladhva, K., Gupta, R. K., Alsaab, H. O., ... & Ratna, R. (2022). A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. Computational Intelligence and Neuroscience: CIN, 2022.

Rani, P., Kumar, R., & Jain, A. (2021). HIOC: a hybrid imputation method to predict missing values in medical datasets. *International Journal of Intelligent Computing and Cybernetics, 14(4),* 598-616.

National Geographic (2022) Rivers and streams. https://education.nationalgeographic.org/resource/resource-library-rivers-and-streams. Accessed 27 December 2022

Vapnik, VN (1998) Statistical learning theory. Adaptive and learning systems for signal processing. *Communications and Control 2*:1-740

Wang, X., Fu, L., & He, C. (2011). Applying support vector regression to water quality modelling by remote sensing data. *International journal of remote sensing, 32(23),* 8615-8627.

Wang, Y., Yuan, Y., Pan, Y., & Fan, Z. (2020). Modeling daily and monthly water quality indicators in a canal using a hybrid wavelet-based support vector regression structure. *Water, 12(5),* 1476.

WHO, UNICEF, World Bank (2022) State of the world's drinking water: an urgent call to action to accelerate progress on ensuring safe drinking water for all. https://www.who.int/publications/i/item/9789240060807. Accessed 02 January 2023

Xie, G., Zhao, Y., Xie, S., Huang, M., & Zhang, Y. (2019). Multi-classification method for determining coastal water quality based on SVM with grid search and KNN. *International Journal of Performability Engineering, 15(10),* 2618.

Xin, L., & Mou, T. (2022). Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification. *Wireless Communications and Mobile Computing,* 2022.

Zhang, X., Yan, C., Gao, C., Malin, B. A., & Chen, Y. (2020). Predicting missing values in medical data via XGBoost regression. *Journal of healthcare informatics research, 4,* 383-394.

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., ... & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health, 1(2*), 107:116.