water

MDPI

Article

Water Quality Prediction Using KNN Imputer and Multilayer Perceptron

Afaq Juna [1,†], Muhammad Umer [1,†], Saima Sadiq [2,†], Hanen Karamti [3], Ala' Abdulmajid Eshmawi [4], Abdullah Mohamed [5] and Imran Ashraf [6,*]

[1] Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[2] Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan
[3] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[4] Department of Cybersecurity, College of Computer Science and Engineering, University of Jeddah, Jeddah 21959, Saudi Arabia
[5] Research Centre, Future University in Egypt, New Cairo 11745, Egypt
[6] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea
* Correspondence: imranashraf@ynu.ac.kr
† These authors contributed equally to this work.

Abstract: The rapid development to accommodate population growth has a detrimental effect on water quality, which is deteriorating. Consequently, water quality prediction has emerged as a topic of great interest during the past decade. Existing water quality prediction approaches lack the desired accuracy. Moreover, the available datasets have missing values, which reduces the performance efficiency of classifiers. This study presents an automatic water quality prediction method that resolves the issue of missing values from the data and obtains a higher water quality prediction accuracy. This study proposes a nine-layer multilayer perceptron (MLP) which is used with a K-nearest neighbor (KNN) imputer to deal with the problem of missing values. Experiments are performed, and performance is compared with seven machine learning algorithms. Performance is further analyzed regarding two scenarios: deleting missing values and the use of a KNN imputer to deal with missing values. Results suggest that the proposed nine-layer MLP model can achieve an accuracy of 0.99 for water quality prediction with the KNN imputer. K-fold cross-validation further corroborates this performance.

Keywords: water quality prediction; KNN imputer; machine learning; multilayer perceptron

Citation: Juna, A.; Umer, M.; Sadiq, S.; Karamti, H.; Eshmawi, A'.A.; Mohamed, A.; Ashraf, I. Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. *Water* **2022**, *14*, 2592. https://doi.org/10.3390/w14172592

Academic Editors: Yulin Tang and Cheng Liu

Received: 16 July 2022
Accepted: 17 August 2022
Published: 23 August 2022

1. Introduction

One of the most crucial natural resources, without which life cannot exist, is water. According to the studies, approximately 66% of the Earth is made up of water with the availability of fresh or usable water being only 1%, while the rest of the water is saline or salt water. Water is an integral part of the prosperity and wealth of a nation. However, the level of water has been falling considerably during the last few decades, which is one of the emerging problems in the modern world. Due to the ever-increasing population of the world, water resources are under pressure to provide basic functions to such a big population due to the water pollution on and under the surface, which is a threatening situation, keeping in mind the depleting water resources. As the population of the world is growing at a rapid pace, it affects its demand, as well as its cost [1]. It can result in the decline of per capita water used at higher population rates. It is proven that a deficiency of clean water can increase the likelihood that people live in poverty. Water is unevenly distributed among the countries. About 60% of the world's water is accessible, which suggests that even though water is abundant on Earth, its accessibility for drinking, agricultural, and commercial use is unevenly distributed geographically [2].

Water quality and abundant water supply are of key importance when it comes to preserving the ecosystem [3]. The availability of clean, fresh water supports social and economic well-being [4]. According to the United Nations Environment Program (2000), 20% of people worldwide lack access to clean drinking water, while around 50% of the population of the world is denied access to safe sanitation systems, which is posing a serious threat to water shortages and waterborne-related diseases. With the increase of approximately 60 million people yearly, an estimated 64 billion m^3 of water is needed to be added annually to the water reservoirs.

The availability of a safe and sufficient quantity of drinking water is a crucial part of basic healthcare since drinking water quality has a significant effect on the health of people. The level of components in potable water must not threaten consumer health or reduce its usefulness [5]. The following characteristics of good water quality should be met by water.

- Free of harmful organisms.
- Clean and clear (low turbidity).
- Lack of saline.
- Devoid of substances that provide an unpleasant flavor or smell.
- Devoid of substances that might have harmful effects on human health.
- Low levels of substances such as lead that are immediately hazardous or have negative long-term effects.
- Free of chemicals that could damage the water supply system or taint washed-in clothing.

1.1. Elements in Water

There are a variety of water quality standards that can be followed depending on the region or country. The World Health Organization (WHO) has created guidelines for the lower and upper limits of several inorganic chemicals that are typically present in drinking water, making them one of the most commonly used standards. The World Health Organization's maximum permissible concentration of elements is described in Table 1.

Table 1. Elements and their admissible amounts for drinking water as mentioned by the WHO [6].

Element	Admissible Amount
Arsenic	10 µg/L
Selenium	40 µg/L
Barium	10 µg/L
Uranium	30 µg/L
Boron	2400 µg/L
Chromium	50 µg/L
Fluoride	1500 µg/L
Organic Species:	
Benzene	10 µg/L
1,4-Dioxane	50 µg/L
Dichloromethane	20 µg/L
Tetrachloroethene	40 µg/L
Nitrilotriacetic acid	200 µg/L
Hexachlorobutadiene	0.6 µg/L
Xylene	500 µg/L
Trichloroethene	20 µg/L
1,2-Dichloroethene	50 µg/L

Table 1. *Cont.*

Element	Admissible Amount
Di(2-ethylhexyl)phthalate	8 µg/L
Pentachlorophenol	9 µg/L
1,4-Dichlorobenzene	300 µg/L
Toluene	700 µg/L
1,2-Dichlorobenzene	1000 µg/L
Carbon tetrachloride	4 µg/L
Ethylbenzene	300 µg/L
Edetic acid	600 µg/L
Styrene	20 µg/L
1,1-Dichloroethane	30 µg/L

1.2. Problem Statement

The public requires access to safe, easily accessible water for drinking, household usage, food production, and recreational activities. A country's economic development may be considerably boosted by better water supply and resource management. Water that is sufficient for personal and household use, constantly available, safe, accessible, and reasonably priced is a right that belongs to everyone. Due to contaminated water every year, lots of people experience kidney failure, cancer, etc. [7]. Contaminated water can lead to diarrhea, cholera, typhoid, etc. [6]. Laboratory methods for water quality classification are time-consuming procedures and need lots of resources. Presently, lots of systems are available for water quality classification but they lack accuracy. Thus, it is the need of the hour to have an automated system that can automatically classify the quality of water with less effort.

1.3. Research Aims and Objectives

Drinking water quality surveillance is the continuous, vigilant evaluation and acceptability of drinking water supplies by the public health community. A perfect distribution system will not keep the public healthy if the water it distributes receives insufficient treatment, and flawless treatment is useless if the system's design or cross-connections allow for contamination. Due to the prevalent problems of water contamination over the past decade, water quality prediction has emerged as a topic of great importance for the survival of life on Earth. Consequently, a large body of automatic water quality prediction approaches can be found in the literature. Predominantly, such works provide comparatively low accuracy. Furthermore, the datasets available for experiments have a large number of missing attributes needed to predict water quality, and the resulting accuracy is low. This study aims to solve these problems by making the following contributions. Firstly, a KNN imputer is used to deal with the missing values problem. Using this technique, the water quality prediction accuracy is distinctly improved compared to deleting the missing values. Secondly, a multilayer perceptron (MLP) is used for water quality prediction. Experiments are performed using different architectures for MLP, including three, six, and nine layers, where the best results are obtained using the customized nine-layered architecture. In addition, a range of machine learning models is used for comparison for the prediction of water quality with and without the KNN imputer.

This study is further divided into four sections. Starting with a description of studies related to current work in Section 2, the proposed methodology, dataset, and machine learning models are presented in Section 3. Section 4 discusses the results while the conclusion is given in Section 5.

2. Related Work

Water is essential to human life since it is used in so many daily activities such as drinking, cooking, maintaining personal cleanliness, farming, and industrial processes [8]. Other processes, such as biotransformation and the creation of electricity, also depend on water [9]. Because human existence depends on the availability of water, both sources (surface and groundwater) are subject to varying degrees of pollution from numerous pollutants [10].

There has been a greater demand for reliable, accurate, and adaptable prediction models as surface water pollution has been acknowledged as a problem and there is growing interest in water quality assessment [11]. Numerous researchers have used neural networks and other machine learning algorithms to forecast water quality in recent years, with promising prediction outcomes [12]. Machine learning models have shown limitations in generalizing complex and highly nonlinear connections between the modeling parameters [13].

Although research has shown that various machine learning models, including deep neural networks, kernel models, fuzzy logic, genetic programming, neuro-inference models, and others, have been utilized to design surface water quality phenomena [14], there are still a number of new classifiers that have not yet been investigated. For the conservation of the water environment, water quality prediction is very important. Authors developed a water quality assessment approach based on long short-term memory (LSTM) and IGRA, taking into account the multivariate correlation and temporal sequence of the water quality data [15]. The first suggestion made by IGRA was to choose features that have a higher absolute correlation being predicted. Second, an LSTM-based prediction model was created, with the indicators collected by IGRA serving as its inputs. Results show promising output for water quality prediction.

Traditional water quality prediction approaches used machine learning and statistical characteristics (normal distribution) and techniques and has not achieved good results. Contrarily, artificial intelligence-based approaches have shown better results as there is no need to determine the relation of dependent variables with independent ones [16]. The authors applied a neural network model to determine the quality of groundwater [17], and provide an improved water quality monitoring system for drinking purposes. Authors investigated quality indicators for potability by applying explainable artificial intelligence [18]. Many researchers have concentrated their efforts on using many variables as a function model after the realization of the significance of monitoring and forecasting the changing water quality. Artificial neural networks (ANNs) such as MLP and radial bias have been employed by researchers for water quality prediction and have achieved satisfactory results [19]. A water quality index was also produced using an ANN and five significant and widely accepted water quality indicators [20]. The literature includes studies on using artificial intelligence systems to predict the water quality index [21].

The recurrent neural network (RNN) is the most widely used deep learning model for the analysis of time-series data. An improved RNN with a significant capacity for information acquisition and archival is LSTM, which has been used extensively for predicting water quality [22,23]. In order to perform a thorough predictive study of the water quality in the next time period, the authors developed a prediction technique based on the bidirectional LSTM that takes into account the reliance at many time scales [24].

For accurate real-time water quality prediction, the researchers introduced a novel feature selection and classification approach in [25]. The complexity of the suggested approach is decreased by using a learning-based model and quantum teaching to choose the best possible collection of characteristics. The authors proposed two tree-based hybrid models, namely, XGBoost and RF, to provide more precise short-term water quality prediction and they also introduced a novel data denoising technique (CEEMDAN) [26].

Numerous versions of models have been utilized by researchers to determine water quality; they are still facing challenges in the process. The literature review indicates that there is a limited number of studies using deep neural networks in predicting water

quality, especially for drinking purposes. In addition, models that can adapt to the complex character of the majority of environmental engineering challenges are required.

3. Materials and Methods

This section presents the proposed approach for water quality prediction, machine learning models, and the dataset used for experiments. The data flow of the proposed approach is depicted in Figure 1. First, the dataset is collected, which contains the electronic health records for different features of water. Since the dataset contains missing values, data preprocessing is needed to deal with this problem. This study leverages the KNN imputer in this regard. Afterward, different machine learning models are applied in addition to the MLP model. For training and testing these models, the data are split into training and testing sets. The classifiers are applied to determine the water quality as potable and not safe for humans.

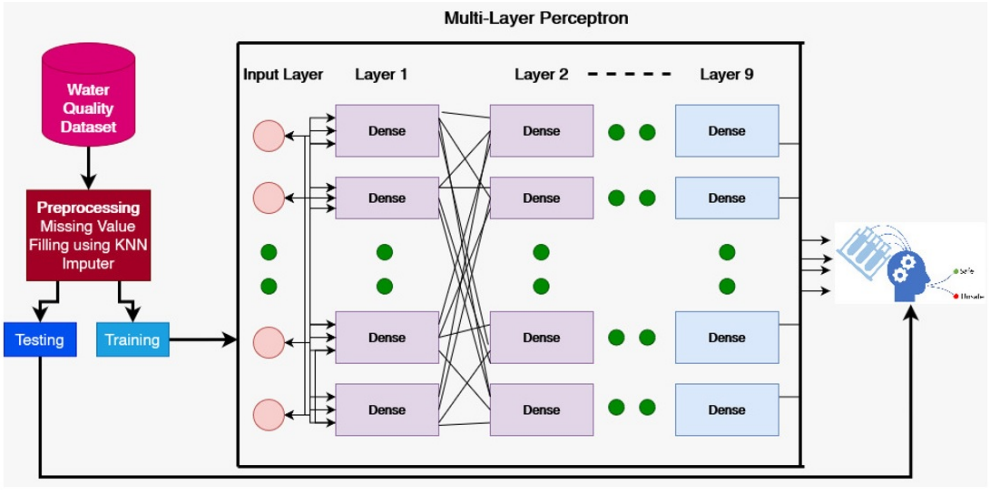


Figure 1. Architecture of the proposed approach.

3.1. Dataset

The dataset utilized in this research is taken from Kaggle. Kaggle is a renowned and free data repository from which we can obtain a dataset without any hassle. The dataset used in this study is available with the name of 'Water Quality' [27]. The dataset consists of 10 columns and it has 935 instances. The target class is potable. It has two values, '0' or '1', where '0' denotes that the water is not safe for drinking and '1' denotes that it is safe for drinking. A detailed dataset description is presented in Table 2.

Table 2. Description of the dataset used in this study.

Feature	Description
pH	pH of water (0 to 14).
Hardness	Capacity of water to precipitate soap in mg/L.
Solids	Total dissolved solids in ppm.
Chloramines	Amount of chloramines in ppm.
Sulfate	Amount of sulfates dissolved in mg/L.
Conductivity	Electrical conductivity of water in μS/cm.
Organic_carbon	Amount of organic carbon in ppm.
Trihalomethanes	Amount of trihalomethanes in μg/L.
Turbidity	Measure of light-emitting property of water in NTU.
Potability	Indicates if water is safe for human consumption. Potable, 1, and not potable, 0.

Data visualization helps researchers find the hidden patterns and relationships among the data attributes [28]. The goal of data visualization is to efficiently and concisely present

facts or information to readers. A chart, map, or infographic is frequently used to visually convey data. Data visualization helps transform data into a more understandable format and shows patterns and outliers. The dataset contains two classes, where 61% of the data belong to the 'not potable' class, and 39% of the data belong to the 'potable' class. The potable class indicates that the water is clean and safe for human use.

Figure 2 shows the histogram distribution of the nine features used for training the machine learning models. The tenth attribute is the target class with potable and not potable values. The histogram helps to understand the distribution of each feature in the dataset. It shows how frequently a value/feature appears in a relatively unbiased way. Figure 2 shows that the given features have a normal distribution and they are not skewed. Moreover, the distribution is unimodal and symmetric. The given range of these features is different, and the occurrence of each feature is helpful to determine the center of a particular feature. For example, Figure 2a shows that the majority of pH values lie between 5.0 and 8.0.

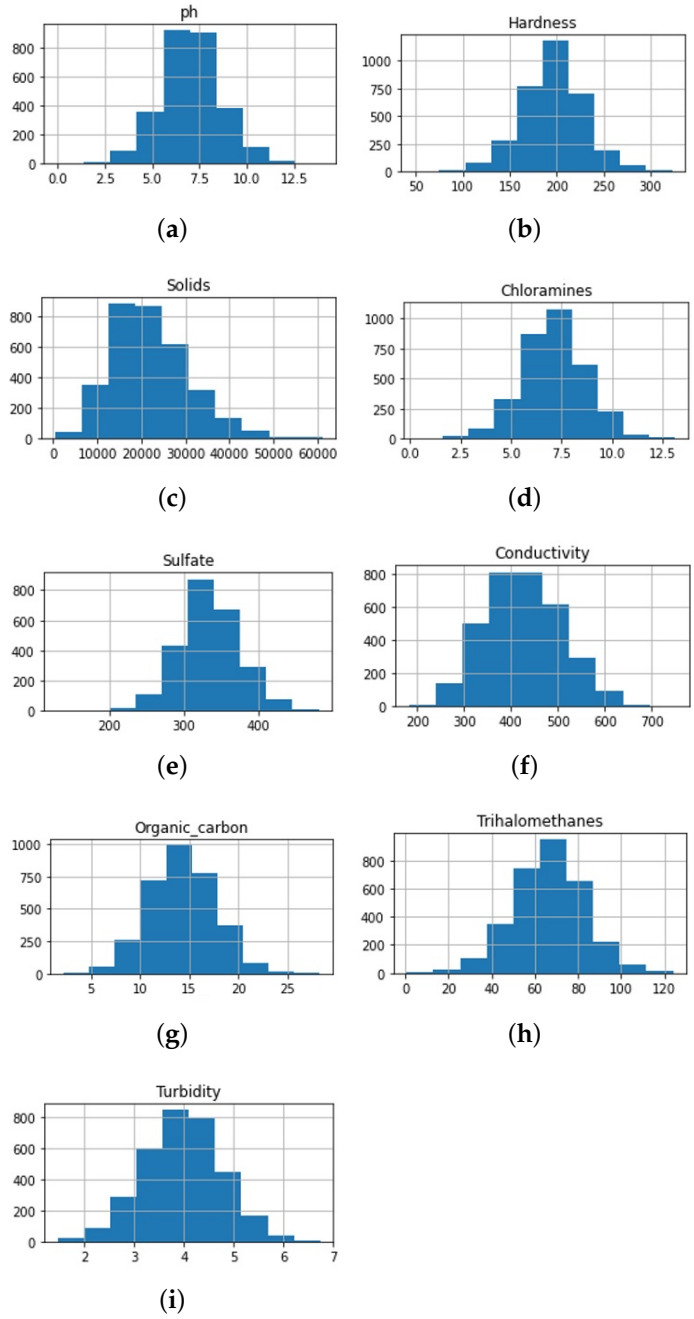


Figure 2. Class-wise histogram representation of each feature.

Dataset attributes have a different level of correlation for water quality prediction. Figure 3 shows the relationship of the features using the heatmap for dataset attributes, which indicates the importance of each attribute with respect to the target class. It indicates that 'solids' and 'chloramines' attributes have a strong linear relationship with potability, while 'organic_carbon' and 'sulfate' have an inverse relationship with potability. This relationship helps to understand which features are important to determine the target class for water quality.

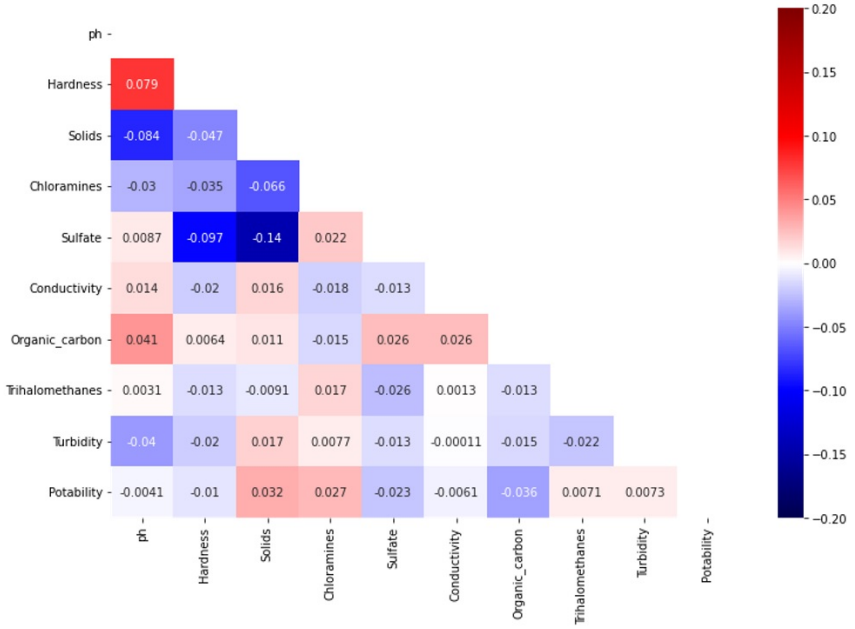


Figure 3. Correlation heatmap of features.

3.2. Data Preprocessing

Data preprocessing is an important step to obtain better performance from the models. In this step, the unnecessary or redundant data are removed from the dataset. These data have no meaning for the machine learning models. Preprocessing helps to enhance the efficacy of the learning models. Not only does preprocessing help to enhance the performance of the model, it also helps to reduce the computational time. In this research, during the data preprocessing, we came to learn that there are several missing values in the dataset. Missing values present in the dataset, according to the class, are presented in Table 3.

Table 3. Missing values in the dataset.

Feature	Missing Values	% of Missing Values
pH	491	14.99
Hardness	0	0.00
Solids	0	0.00
Chloramines	0	0.00
Sulfate	781	23.84
Conductivity	0	0.00
Organic_carbon	0	0.00
Trihalomethanes	162	4.95
Turbidity	0	0.00
Potability	0	0.00

It is clear from the table that there are a large number of missing values. As the dataset is categorical, the missing values can be handled with two different methods:

- Using KNN imputer.
- By removing the missing value from the dataset.

3.2.1. KNN Imputer

In the modern world, data are collected from various sources and used for analysis, insight generation, theory validation, and other purposes. There may frequently be some information missing from these data which is gathered from various sources. This can be the result of an issue with data extraction or collection caused by human error. Thus, dealing with these missing values becomes a crucial step in the preprocessing of data. The choice of imputation method is important because it can have a big impact on the models' performance. A popular technique for imputing missing values is KNN imputer by sci-kit-learn. It is frequently used as an alternative to conventional imputation methods [29]. By using the Euclidean distance matrix to find the nearest neighbors, the KNN imputer aids in the impute of missing values that are present in the observations. By ignoring the missing values and increasing the weight of the non-missing coordinates, the Euclidean distance is determined. Mathematically, Euclidean distance can be calculated as:

$$D_{xy} = \sqrt{weight * squared\ distance\ from\ present\ coordinates} \quad (1)$$

where

$$weight = \frac{total\ number\ of\ coordinates}{number\ of\ present\ coordinates} \quad (2)$$

3.2.2. Removing Missing Values from Dataset

The second option for handling the data is to remove the missing values from the data. The second set of experiments is performed using this approach where all the fields with the missing values are removed.

3.3. Machine Learning Models Used in Study

Machine learning plays a significant role in enhancing the accuracy and efficacy of water quality classification. To classify water quality, there exists a variety of machine learning algorithms. The Scikit-learn library of Python has a variety of machine learning classifiers. This library is open source and has a sizeable user base; it largely contributes to the research community. This study uses the Scikit-learn library to implement LR, SVC, DT, RF, KNN, SGDC, and XGBoost.

3.3.1. Logistic Regression

For solving the binary classification problem, LR is one of the widely used methods. The logistic equation, often known as the sigmoid function, is the technique that makes LR so popular [30]. Any evaluated number may be given to the sigmoid function (S-shaped curve) that converts it to a number between 0 and 1.

$$y = \frac{1}{(1 + e^{-value})} \quad (3)$$

where e represents the base of algorithms. The real numerical values are to be transformed. The logistic functions value also ranges between 0 and 1.

$$y = \frac{e^{b_0 + b_1 * x}}{(1 + e^{b_0 + b_1 * x})} \quad (4)$$

where y presents the expected performance, b_0 is the bias or intercept, and b_1 presents the coefficient for x which is an input value.

In the input data, each column has a coefficient that is correlated with the training data. To achieve a higher value of accuracy, LR uses 'max_iter' for solving converge.

3.3.2. Support Vector Classifier

SVC is also known as a support vector machine. It is a famous supervised machine learning algorithm. SVC performs classification problems by developing the best line in two dimensions [31]. SVC uses the RBC kernel and can be used to find the regression line. The general equation is:

$$k(a_a - a_2) = exp\left(\frac{||(a_1 - a_2)||^2}{2a^2}\right) \tag{5}$$

where k is the kernel function, and $(a_1 - a_2)$ is the distance between a_1 and a_2. The kernel function k can be written as

$$k = \frac{1}{e^{\frac{d_{12}^2}{2\sigma^2}}} \tag{6}$$

where σ is the hyperparameter.

3.3.3. Decision Tree

DT is a renowned machine learning algorithm extensively utilized for regression and classification problems. The selection of the root node at each level is a problem in the decision tree [32]. This process is termed 'attribute selection'. For attribute selection, there are two renowned techniques: the 'Gini index' and 'information gain'. The Gini index can be computed by the following equation:

$$Gini = 1 - \sum_{i=1}^{classes} p(\frac{i}{t})^2 \tag{7}$$

The Gini index helps to compute the impurity of data in the dataset. Another attribute selection technique is information gain. Information gain tells about the purity of data. Information gain can be calculated when we have the entropy of the target class and entropy of the each attribute. Entropy D can be calculated as:

$$entropy(D) = -\sum_{i=1}^{|c|} P_r(c_i) log_2 P_r(c_i) \tag{8}$$

$$\sum_{i=1}^{|c|} P_r(c_i) = 1 \tag{9}$$

where $P_r(c_i)$ presents the probability, c_i presents the class, and D presents the dataset.

The entropy of attribute A_i is utilized as the current root and can be calculated as:

$$entropyA_i(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} * entropy(D_j) \tag{10}$$

Finally, the following information is gained when attribute Ai is chosen to branch or split data:

$$entropy(D, A_i) = entropy(D) - entropyA_i(D) \tag{11}$$

3.3.4. Random Forest

A tree-based classifier RF combines several poor apprentices (poor learners) to generate very accurate predictions. To train different decision trees utilizing diverse bootstrap samples, RF uses bootstrap bagging [33]. A bootstrap sample is generated using the sub-sampling of the training dataset, where the size of the training and test sample dataset is the

same. Similar to other ensemble classifiers, RF uses decision trees for making predictions. At each stage, the identification of the root node is a challenging task for the development of decision trees.

$$p = mode\{T_1(y), T_2(y), T_3(y), \ldots\ldots\ldots\ldots, T_m(y)\} \tag{12}$$

$$p = mode\{\sum_{m=1}^{m} T_m(y)\} \tag{13}$$

The number of decision trees participating in the prediction process is $T_1(y)$, $T_2(y)$, $T_3(y)$, . . . , $T_m(y)$, and p is the decision made by the decision trees by a majority vote.

In RF, the term "random state" is used during training to regulate the unpredictability of the sample.

3.3.5. K-Nearest Neighbor

KNN is a simple supervised classification technique. The KNN locates the similarity between the previous examples and new data and then places the new data in the group with high similarity [34]. Distance calculations between the existing samples and new data are used to determine similarity. Different distance estimation techniques, such as Euclidean, Manhattan, and Minkowski, are used for measuring distance. The KNN technique can be applied to classification and regression problems. Due to the fact that KNN is a nonparametric method, it does not evaluate any inferences about the underlying data.

3.3.6. Stochastic Gradient Decent Classifier

The working principle of SGDC relies on the working of LR and SVM. SGDC uses the LR convex loss function and proves to be a powerful classifier. It is an excellent option for multiclass categorization. SGDC aggregates multiple classifiers in the OvA (one-versus-all) method [35]. The quality of SGDC is that it handles large datasets efficiently. It uses a single example per iteration. As SGDC uses the regression technique, it is very easy to implement and easy to understand. For better results, SGDC must be correctly valued. SGDC has a high value of sensitivity in terms of feature scaling.

3.3.7. XGBoost

XGBoost is a fast supervised learning algorithm. For the accurate and precise classification of water quality, XGBoost is used in this study. Due to the availability of regularized learning features, it helps in smoothing the final weights and it avoids the overfitting phenomenon [36]. The specific algorithm is as follows:

$$\Omega(\theta) = \sum_{i=1}^{n} d(y_i, \hat{y}_i) + \sum_{k=1}^{k} \beta(f_k) \tag{14}$$

where d is the loss function, β is the regularization term, y_i is the predicting value, n is the instance number in training, and k is the number of trees.

3.4. Deep Learning Models Used in Study

Due to promising results and high accuracy values, deep neural networks are in the eye of many researchers. Water quality classification is also performed using the deep learning model MLP in this study. A brief description of MLP is given in this section.

Multilayer Perceptron

An input vector and the associated output vector are nonlinearly mapped by the deep learning neural network known as MLP. MLP constitutes a hidden layer, input layer, and output layer. For the activation of neurons, MLP uses nonlinear activation function except for the input node [37]. Due to the availability of a nonlinear function for activation, MLP can handle data that cannot be linearly separated [38]. The connection weights are

modified, and the calculation takes into account both the desired and actual output. The output node is presented by y at the nth data point. The error can be computed by the following equation:

$$e_y(n) = t_y(n) - p_y(n) \tag{15}$$

where p is the output of perceptron and t presents the target value.

Node weights are changed following the adjustment to lessen the inaccuracy of the overall output.

$$\in (n) = \frac{1}{2}\sum_{y} e2y(n) \tag{16}$$

Any change in the weight is given using gradient descent.

$$\Delta w_{yx}(n) = \eta \frac{\delta \epsilon(n)}{(\delta v_y(n))} p_x(n) \tag{17}$$

where η represents the learning rate and p_x is the output of the previous neuron.

Derivation calculation is calculated with the help of vy, which is an induced local field. This derivative is calculated as

$$\frac{\delta \epsilon(n)}{(\delta v_y(n))} = e_y(n)\phi'(v_y(n)) \tag{18}$$

ϕ is a derivative of the constant activation function.

$$\frac{\delta \epsilon(n)}{(\delta v_y(n))} = e_y(n)\sum_{k} \frac{\delta \epsilon(n)}{(\delta v_y(n))} w_{ky}(n) \tag{19}$$

3.5. Proposed Approach for Water Quality Prediction

The dataset for this study was taken from Kaggle, a well-known data source. After obtaining the dataset, the preprocessing was carried out to resolve the problem of missing values. These missing values had a strong impact on the efficacy of the learning models. To handle the missing values, KNN imputer was used. After that, the data splitting was performed in 70:30, with 70% for the training of the model and 30% of the data used for testing purposes. The proposed MLP system is used for water quality classification. It is a feedforward deep learning network that gives a mapping between the matching output vector and an input vector. The proposed MLP system consists of nine layers with each layer of 250 neurons. MLP is used with the "binary_crossentropy" loss function while "Adam" is used as the optimizer. The learning rate is set to 0.001, while a batch size of 100 is used with 20 epochs for training the model.

4. Results and Discussion

The results of water quality classification using various classifiers are discussed in this section. Machine and deep learning models were employed using Python 3.0 on a Jupyter notebook. Experiments were carried out on the Core i7 7th-generation machine with Windows 10 as the operating system. The accuracy, precision, recall, and F1 score of the learning models are used to evaluate their performance.

4.1. Results of Machine Learning Models with Deleted Missing Values

In the first set of experiments, the missing values are deleted from the dataset. After deleting the missing values from the dataset, machine learning models are applied to the data. The results of the machine learning models obtainedby deleting missing values from the dataset are presented in Table 4.

Table 4. Results of models obtainedby deleting missing values from the dataset.

Model	Accuracy	Precision	Recall	F1 Score
LR	0.48	0.48	0.48	0.48
SVC	0.52	0.54	0.52	0.47
DT	0.72	0.72	0.72	0.72
RF	0.79	0.79	0.79	0.79
KNN	0.57	0.55	0.57	0.56
SGDC	0.50	0.25	0.50	0.33
XGBoost	0.76	0.76	0.76	0.76

From the results, it is clear that the RF and the XGBoost achieve accuracy values of 79% and 76%, respectively, which are the highest among all models. RF achieves a 79% value for the precision, recall, and F1 score, and XGBoost achieves a 76% value for the precision, recall, and F1 score. LR shows the worst performance and achieves 48% each for accuracy, precision, recall, and F1 score. Overall, the performance results of machine learning models are not satisfactory while using the deleted missing value data. A graphical representation of the machine learning model results with deleted missing values data is given in Figure 4. It indicates that besides the performance of RF and XGBoost, the results from other models are very poor and unsatisfactory.

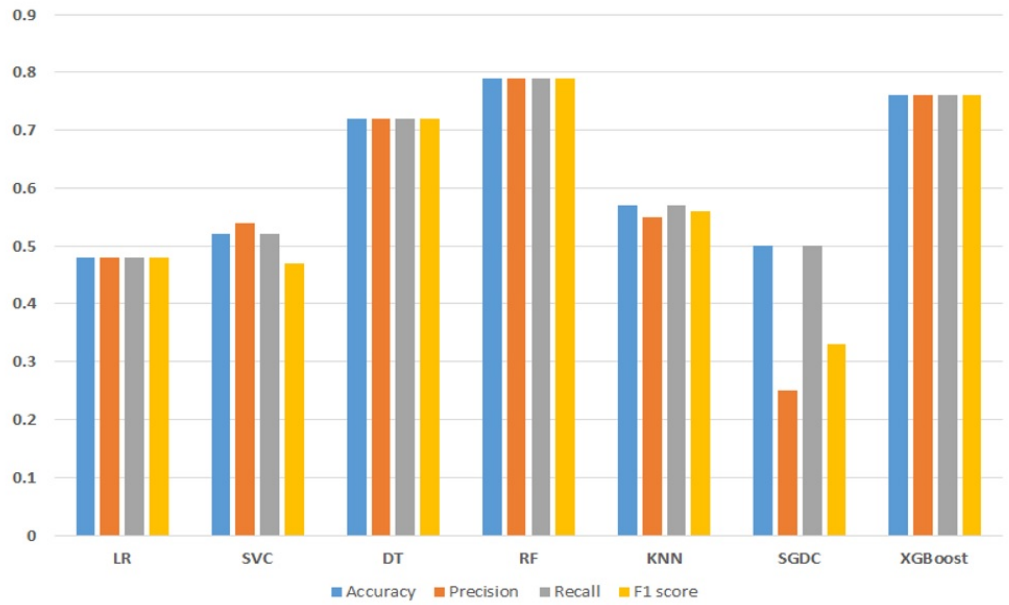


Figure 4. Graphical representation of machine learning models results using the deleted missing values data.

In addition, Figure 5 shows the standard deviation and mean absolute error for the models used in this study. The standard deviation is calculated using 10-fold cross-validation. It shows that RFC and XGB have the lowest mean absolute error while DT has the lowest standard deviation.

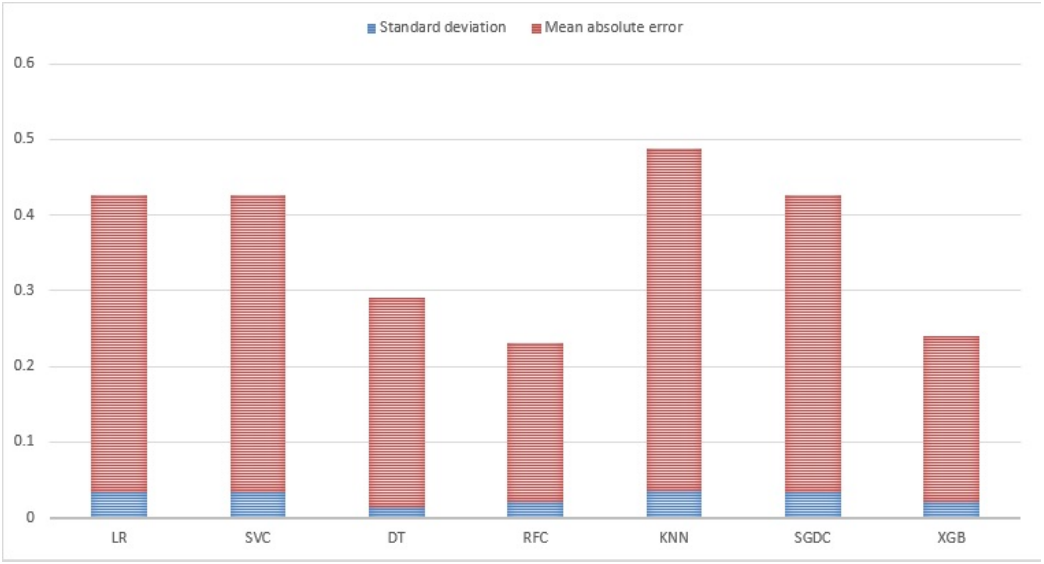


Figure 5. Standard deviation and mean absolute error of models.

4.2. Results of Machine Learning Models by Using KNN Imputer

The second set of experiments is performed using the KNN imputer. After the preprocessing of the dataset, several missing values are found in the dataset. To handle the missing values, we used the KNN imputer. The KNN imputer imputes the value based on the mean of the given values using the Euclidean distance. Once the missing values are imputed, the data are used for experiments with the machine learning models. The results of the machine learning models using the KNN imputer are shown in Table 5.

Table 5. Results of machine learning models using KNN imputer.

Model	Accuracy	Precision	Recall	F1 Score
LR	0.61	0.38	0.61	0.47
SVC	0.61	0.38	0.61	0.47
DT	0.72	0.73	0.72	0.72
RF	0.80	0.80	0.80	0.80
KNN	0.59	0.59	0.59	0.58
SGDC	0.59	0.56	0.59	0.55
XGBoost	0.80	0.80	0.80	0.79

The confusion matrices for all the models are provided in Figure 6. It can be observed that DT achieved the highest number of true positives, with 249 correct predictions, and XGBoost achieved the highest true negatives, with 529 correct predictions, from the used machine learning models. KNN shows the highest type I and type II errors, with 244 false positives and 157 false negatives. The MLP has no wrong predictions for the given dataset.

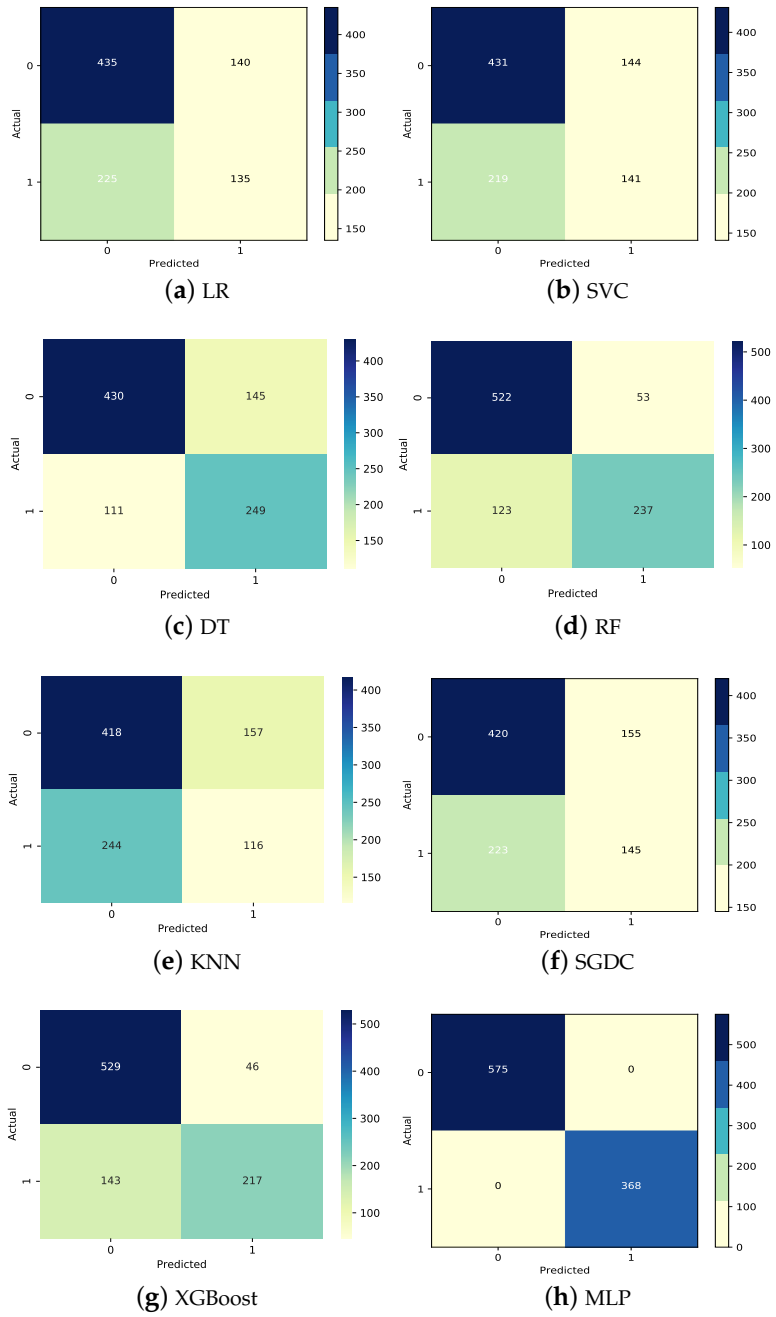


Figure 6. Confusion matrix of machine learning models.

From the results, it is clear that the RF and the XGBoost achieve accuracy values of 80% while RF achieves an 80% precision, recall, and F1 score. XGBoost achieves 80% each for precision and recall but its F1 score is 79%. KNN and SGDC achieve the lowest accuracy value of 59%. A graphical representation of the machine learning model results using the KNN imputer is provided in Figure 7. It demonstrates that using the KNN imputer improves the machine learning model performance.

The area under the curve (AUC) is a metric used to assess a classifier's capacity to discriminate between classes. The performance of the model in differentiating between the positive and negative classes improves with the increasing AUC. AUC curve of the machine learning models is presented in Figure 8. It is evident that AUC for XGBoost and RF is higher than that of other models.

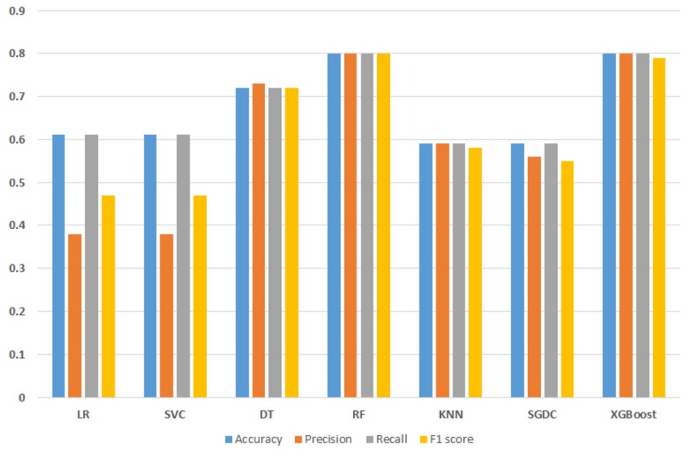
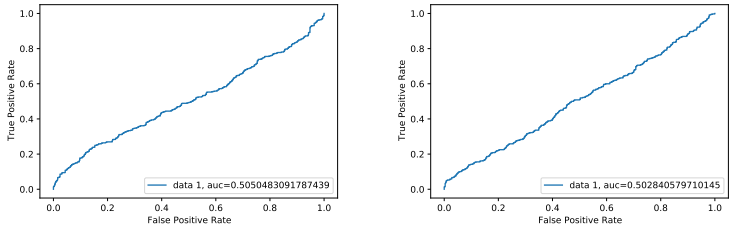
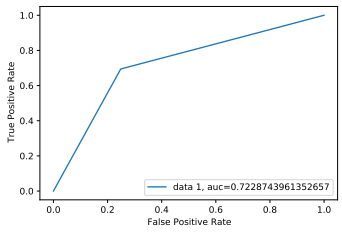


Figure 7. Results of machine learning models using KNN imputer to handle missing values.

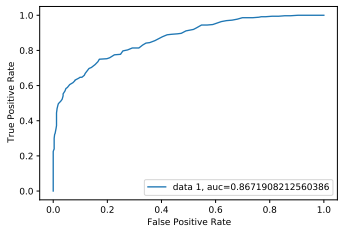


(**a**) LR

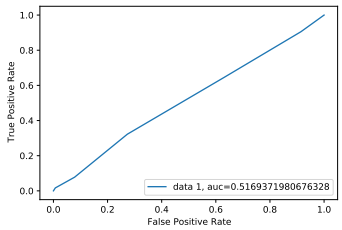
(**b**) SVC



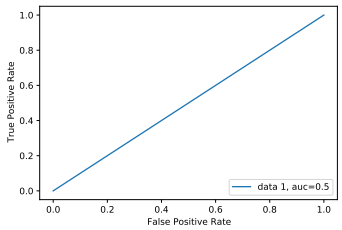
(**c**) DT



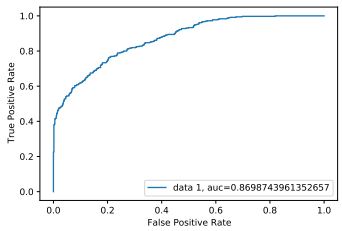
(**d**) RF



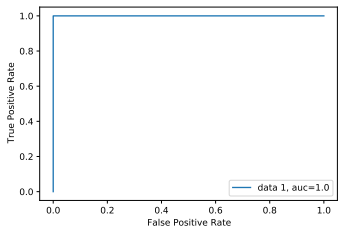
(**e**) KNN



(**f**) SGDC



(**g**) XGBoost



(**h**) MLP

Figure 8. AUC curve of models.

4.3. Comparison of Machine Learning Models with and without KNN Imputer

For clarity and performance analysis, we compare the results of the machine learning models with and without using the KNN imputer. The comparison reveals that the performance of the machine learning models in the second experiment (using the KNN imputer) is best, as compared to the results achieved by the learning models using the deleted missing values data. The results of the machine learning models for both scenarios are given in Table 6.

Table 6. Accuracy comparison of the machine learning models.

Model	Accuracy	
	KNN Imputer	Deleting Missing Values
LR	0.61	0.48
SVC	0.61	0.52
DT	0.72	0.72
RF	0.80	0.79
KNN	0.59	0.57
SGDC	0.59	0.50
XGBoost	0.80	0.76

Figure 9 portrays the difference in the performance of machine learning models when used with deleted missing values and using the imputed dataset with the KNN imputer. The KNN imputer not only improves the individual performance of the models but also leads to overall better performance from all the machine learning models.

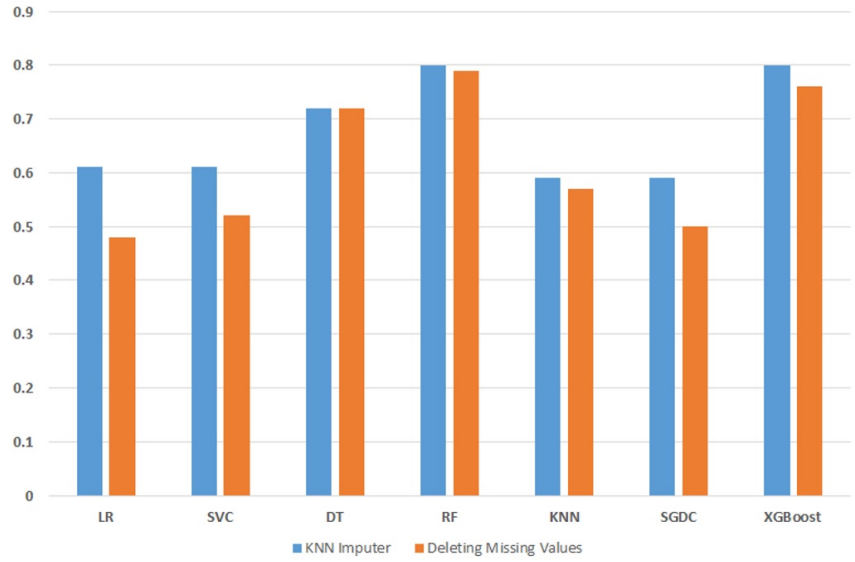


Figure 9. Graphical representation of machine learning models results using KNN imputer to handle missing values.

4.4. Results of Proposed Multilayer Perceptron

We experimented with three different architectures of the proposed MLP. The basic difference lies in the fact that a different number of layers are used with each architecture, starting with three layers and increasing to six and nine later on. Similar to the machine learning process, MLP is applied with deleted missing values and the KNN imputer. Table 7 shows the results with each architecture of MLP when missing values are deleted. Although there is improvement in the performance of MLP after increasing the number of layers from three to nine, the performance is poor, with the best accuracy of 0.75, when used with nine-layer architecture.

Table 7. Results using the multilayer perception with 3, 6, and 9 layers with deleted missing value data.

Model	Accuracy	Precision	Recall	F1 Score
MLP-3	0.6545	0.71	0.85	0.78
MLP-6	0.6932	0.75	0.87	0.82
MLP-9	0.7532	0.78	0.89	0.85

Table 8 shows the results of the MLP when KNN imputed data are used for training. Results indicate that when using three layers, the performance is poor, with an 0.8271 accuracy score. However, as the number of layers is increased from three to six, the accuracy is increased to 0.91. From the results, it is clear that the MLP achieves 99.90% accuracy, precision, recall, and F1 score when used with a nine-layer architecture.

Table 8. Results using the multilayer perception with three, six, and nine layers using KNN imputer.

Model	Accuracy	Precision	Recall	F1 Score
MLP-3	0.8271	0.78	0.82	0.80
MLP-6	0.9124	0.87	0.91	0.90
MLP-9	0.9990	0.993	0.991	0.993

4.5. Comparison of Machine Learning Models with Proposed MLP

Table 9 presents the performance comparison of machine learning models and the proposed approach for both scenarios. Results reveal that both the proposed approach and the machine learning models perform better when the KNN imputer is used to fill the missing values in the water quality prediction data. However, the best results are obtained using the proposed approach, which obtains 100% accuracy when used with the KNN imputer.

Table 9. Accuracy of all models using KNN imputer.

Scenario	Accuracy									
	LR	SVC	DT	RF	KNN	SGDC	XGBoost	MLP 3 Layer	MLP 6 Layer	MLP 9 Layer
KNN imputer	0.61	0.61	0.72	0.80	0.59	0.59	0.80	0.82	0.91	0.997
Deleting MV	0.48	0.52	0.72	0.79	0.57	0.50	0.76	0.65	0.69	0.75

4.6. Results Using K-Fold Cross-Validation

To further validate the results of the proposed approach, we perform a 10-fold cross-validation using the proposed MLP model with the KNN imputer-filled dataset. Table 10 presents the results of each fold for the proposed model. The results show a 100% accuracy for each fold with slight variations in the precision, recall, and F1 score.

Table 10. Ten-fold cross-validation classification results using MLP-9.

Fold Number	Accuracy	Precision	Recall	F1 Score
1st-Fold	1.00	0.994	0.994	0.994
2nd-Fold	0.998	1.000	1.000	1.000
3rd-Fold	1.00	0.997	0.999	0.998
4th-Fold	0.996	1.000	1.000	1.000
5th-Fold	0.999	1.000	1.000	1.000
6th-Fold	0.998	1.000	1.000	1.000
7th-Fold	1.00	1.000	1.000	1.000
8th-Fold	1.00	1.000	1.000	1.000
9th-Fold	1.00	1.000	1.000	1.000
10th-Fold	0.999	1.000	1.000	1.000
Average	**0.999**	**0.9991**	**0.9993**	**0.9992**

5. Conclusions

Recent environment-affecting developments have led to water contamination which has negativeeffects on human life and causes several complicated diseases. Consequently, water quality prediction is essential for sustaining human life. This study performed experiments regarding the water quality prediction where the dataset has missing values. For obtaining highly accurate predictions, MLP is used with the KNN imputer to deal with missing values. Extensive experiments were carried out using several machine learning models with two different scenarios: deleting missing values and using the KNN imputer. Results suggest that the use of the KNN imputer for filling the missing values is a better choice and it produces better results. The MLP obtains the best accuracy of 99.9% with a nine-layer architecture and the KNN imputer. Overfitting is the main problem faced by machine learning models with imbalanced datasets. In the future, using mixed features and a balanced dataset is intended to obtain generalized results. We also intend to use deep learning with a large dataset for automatic feature extraction and water quality prediction.

Author Contributions: Conceptualization, M.U. and S.S.; data curation, A.J. and S.S.; formal analysis, M.U.; funding acquisition, A.M.; investigation, A'.A.E.; methodology, H.K. and I.A.; project administration, A'.A.E.; resources, A.M.; software, H.K. and A'.A.E.; supervision, I.A.; validation, A.M.; visualization, H.K.; writing—original draft, A.J., M.U. and S.S.; writing—review and editing, I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available on Kaggle and can be requested from all co-authors.

Acknowledgments: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2022R192), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Muhammad, S.Y.; Makhtar, M.; Rozaimee, A.; Aziz, A.A.; Jamal, A.A. Classification model for water quality using machine learning techniques. *Int. J. Softw. Eng. Its Appl.* **2015**, *9*, 45–52. [CrossRef]
2. Radhakrishnan, N.; Pillai, A.S. Comparison of water quality classification models using machine learning. In Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 1183–1188.
3. Walley, W.; Džeroski, S. Biological monitoring: A comparison between Bayesian, neural and machine learning methods of water quality classification. In *Environmental Software Systems*; Springer: Berlin/Heidelberg, Germany, 1996; pp. 229–240.
4. Nasir, N.; Kansal, A.; Alshaltone, O.; Barneih, F.; Sameer, M.; Shanableh, A.; Al-Shamma'a, A. Water quality classification using machine learning algorithms. *J. Water Process. Eng.* **2022**, *48*, 102920. [CrossRef]
5. Sillberg, C.V.; Kullavanijaya, P.; Chavalparit, O. Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River. *J. Ecol. Eng.* **2021**, *22*, 70–86. [CrossRef]
6. World Health Organization. Drinking water. Available online: https://www.who.int/news-room/fact-sheets/detail/drinking-water#:~:text=Contaminated%20water%20and%20poor%20sanitation,individuals%20to%20preventable%20health%20risks (accessed on 5 May 2022).
7. Abdulla, A.F. Bacterial Diseases Caused By Contaminated Drinking Water. 2021 .
8. Luo, Z.; Shao, Q.; Zuo, Q.; Cui, Y. Impact of land use and urbanization on river water quality and ecology in a dam dominated basin. *J. Hydrol.* **2020**, *584*, 124655. [CrossRef]
9. Okumah, M.; Yeboah, A.S.; Bonyah, S.K. What matters most? Stakeholders' perceptions of river water quality. *Land Use Policy* **2020**, *99*, 104824. [CrossRef]
10. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **2021**, *9*, 104599. [CrossRef]
11. Abba, S.; Hadi, S.J.; Sammen, S.S.; Salih, S.Q.; Abdulkadir, R.; Pham, Q.B.; Yaseen, Z.M. Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *J. Hydrol.* **2020**, *587*, 124974. [CrossRef]
12. Rajaee, T.; Khani, S.; Ravansalar, M. Artificial intelligence-based single and hybrid models for prediction of water quality in rivers: A review. *Chemom. Intell. Lab. Syst.* **2020**, *200*, 103978. [CrossRef]

13. Barzegar, R.; Adamowski, J.; Moghaddam, A.A. Application of wavelet-artificial intelligence hybrid models for water quality prediction: A case study in Aji-Chay River, Iran. *Stoch. Environ. Res. Risk Assess.* **2016**, *30*, 1797–1819. [CrossRef]
14. Tiyasha; Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670. [CrossRef]
15. Zhou, J.; Wang, Y.; Xiao, F.; Wang, Y.; Sun, L. Water quality prediction method based on IGRA and LSTM. *Water* **2018**, *10*, 1148. [CrossRef]
16. Wagh, V.; Panaskar, D.; Muley, A.; Mukate, S.; Gaikwad, S. Neural network modelling for nitrate concentration in groundwater of Kadava River basin, Nashik, Maharashtra, India. *Groundw. Sustain. Dev.* **2018**, *7*, 436–445. [CrossRef]
17. Bilali, A.E.; Taleb, A.; Mazigh, N.; Mokhliss, M. Prediction of chemical water quality used for drinking purposes based on artificial neural networks. *Moroc. J. Chem.* **2020**, *8*, 8–3.
18. Dwivedi, P.; Khan, A.A.; Mudge, S.; Sharma, G. Explainable AI (XAI) for Social Good: Leveraging AutoML to Assess and Analyze Vital Potable Water Quality Indicators. In *Computational Intelligence in Data Mining*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 591–606.
19. Ubah, J.; Orakwe, L.; Ogbu, K.; Awu, J.; Ahaneku, I.; Chukwuma, E. Forecasting water quality parameters using artificial neural network for irrigation purposes. *Sci. Rep.* **2021**, *11*, 24438. [CrossRef]
20. Gupta, R.; Singh, A.; Singhal, A. Application of ANN for water quality index. *Int. J. Mach. Learn. Comput* **2019**, *9*, 688–693. [CrossRef]
21. Aldhyani, T.H.; Al-Yaari, M.; Alkahtani, H.; Maashi, M. Water quality prediction using artificial intelligence algorithms. *Appl. Bionics Biomech.* **2020**, *2020*, 6659314. [CrossRef]
22. Liu, P.; Wang, J.; Sangaiah, A.K.; Xie, Y.; Yin, X. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. *Sustainability* **2019**, *11*, 2058. [CrossRef]
23. Hu, Z.; Zhang, Y.; Zhao, Y.; Xie, M.; Zhong, J.; Tu, Z.; Liu, J. A water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. *Sensors* **2019**, *19*, 1420. [CrossRef]
24. Zou, Q.; Xiong, Q.; Li, Q.; Yi, H.; Yu, Y.; Wu, C. A water quality prediction method based on the multi-time scale bidirectional long short-term memory network. *Environ. Sci. Pollut. Res.* **2020**, *27*, 16853–16864. [CrossRef]
25. Charles, J.; Vinodhini, G.; Nagarajan, R. An efficient feature selection with weighted extreme learning machine for water quality prediction and classification model. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 1969–1994.
26. Lu, H.; Ma, X. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* **2020**, *249*, 126169. [CrossRef] [PubMed]
27. Kaggle. Water Quality Available online: https://www.kaggle.com/datasets/adityakadiwal/water-potability (accessed on 5 May 2022).
28. Aparicio, M.; Costa, C.J. Data visualization. *Commun. Des. Q. Rev.* **2015**, *3*, 7–11. [CrossRef]
29. Saranya, N.; Samyuktha, M.S.; Isaac, S.; Subhanki, B. Diagnosing chronic kidney disease using KNN algorithm. In Proceedings of the 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 19–20 March 2021; Volume 1, pp. 2038–2041.
30. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
31. Wien, M.; Schwarz, H.; Oelbaum, T. Performance analysis of SVC. *IEEE Trans. Circuits Syst. Video Technol.* **2007**, *17*, 1194–1203. [CrossRef]
32. Chandra, B.; Varghese, P.P. Fuzzy SLIQ decision tree algorithm. *IEEE Trans. Syst. Man, Cybern. Part* **2008**, *38*, 1294–1301. [CrossRef] [PubMed]
33. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote. Sens.* **2016**, *114*, 24–31. [CrossRef]
34. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
35. Sowmya, B.; Nikhil Jain, C.; Seema, S.; KG, S. Fake News Detection using LSTM Neural Network Augmented with SGD Classifier. *Solid State Technol.* **2020**, *63*, 6985–9665.
36. CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system; In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016.
37. Arber, S.; Hunter, J.J.; Ross Jr, J.; Hongo, M.; Sansig, G.; Borg, J.; Perriard, J.C.; Chien, K.R.; Caroni, P. MLP-deficient mice exhibit a disruption of cardiac cytoarchitectural organization, dilated cardiomyopathy, and heart failure. *Cell* **1997**, *88*, 393–403. [CrossRef]
38. Zhang, C.; Pan, X.; Li, H.; Gardiner, A.; Sargent, I.; Hare, J.; Atkinson, P.M. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *140*, 133–144. [CrossRef]