2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Oct 13-14, 2022

2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-7433-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRITO56286.2022.9964809

Mean and Prediction Imputation-Based Approach for Predicting Water Potability Using Machine Learning

[1]Saqib Alam Ansari
Maharishi University of Information Technology, India
upGrad Education Private Limited, India
saqibdatascience@gmail.com

[2]Chetan Sharma*
upGrad Education Private Limited, India
chetanshekhu@gmail.com

[3]Trapty Agarwal
Maharishi University of Information Technology,
India
trapty@gmail.com

***Abstract* -Water is one of the most critical components of our universe, as it is what keeps us alive. Therefore, the portability of a person's water supply is essential to their survival. Using the water potability indicator, you can determine whether or not the water in question is safe for human consumption as drinking water. Water's ability to support human consumption is inversely related to its concentration of minerals, particularly pH, sulfate, and chloramines, as is well-known; in this study, they took a water portability dataset from Kaggle publicly available. Unfortunately, every dataset includes some null values. Therefore, two approaches are proposed in this study: mean imputation and predicting imputation for replacing the null values. Six machine learning algorithms with each approach are implemented, achieving an accuracy of 66.23% and 70.09%. In the future, the Prediction imputation approach can be implemented on a different dataset.**

Keywords: Water Portability, Machine Learning, Mean Imputation, Prediction Imputation, ANN

I. INTRODUCTION

Water is one of the essential things responsible for life on earth and the existence of living beings, and life cannot be imagined without water [1]. A human body contains up to 60 % water, as no living being can survive for extended periods without drinking water. In contrast, some organisms, such as viruses and bacteria, keep their cells filled with water, and some have 90 % of their weight from water [2]. Also, no life can be imagined without it. It provides essential nutrients and minerals to the body. Water is more than a necessity for life. It has the potential for multiple regions on Earth—in the form of an indispensable commodity, potable water. However, numerous parameters determine whether the water is drinkable, including the users' safety factors. However, multiple parameters determine whether the water is potable, including hardness, solids, ph, sulfate, etc. Suppose we discuss the latest statistics about the volume of water present on this globe. In that case, only 0.5% of water is drinkable and fresh, where 99% of freshwater is below the ground surface, and every 1 of the ten people lacks water. The water resources in India are about only 4% of the world, with 18% of the world's population, making it a water-stressed country [3]. The water makes it not potable because some substances present in it may harm the body that consumes it. Some of these substances may not break if present below a specific range, but excess of them make water unsafe for drinking. Some of them are impurities whose presence is hazardous to water. Discussing some contaminants like soil, dunk, and stone may be removed by some filtrations, but at the microscopic level, some solutes are present whose concentration should be in limit. But post-measuring those solutes in water may tell us about the potability of water. In this research, we proposed machine learning algorithms that can be used to check water potability by using some parameters present in water. Our factors for determining water potability depend upon pH level, hardness, solids, chloramines, trihalomethanes, turbidity, sulfate, conductivity, and organic carbon measurement. The potability of water depends on the amount of these factors present in water. One can trust this algorithm's result of potability as this algorithm have 70% of accuracy.

II. STATE OF THE ART

The decision tree and KNN model are used to find water potability and the dataset used in this research is the Kaggle dataset [4]. The accuracy score of the Decision Tree model is 58.5, and the KNN model is 61.7 % [5]. The dataset used in this research is the Kaggle dataset [4]. Models used in research are Random forest with 82.45% accuracy, Deep Neural Network with 84.57% accuracy, Gaussian Naïve Bayes with 92.44 % accuracy, Artificial Neural Network with 98.12 % accuracy, and Support Vector Machine with 93.17 % of accuracy [6]. The data were collected by the USGS website [7]. Different algorithms used for individual features with average mean absolute percentage error for the algorithms LSTM, LSSVM, RBFNN, PSO-SVM, CEEMDAN-XGBoost, CEEMDAN-RF with 7.35%, 6.96%, 6.93%, 8.96%, 3.71%, 3.90%, 4.60% and 6.41% respectively [8]. The data used in the study was collected from Laguna Lake Development Authority (LLDA). This study uses Decision trees, Naïve Bayes, Random Forest, Gradient Boost, and Deep learning algorithms with a classification accuracy of 87.69%, 72.31%, 78.46%, 73.80%, and 72.30%, respectively [9]. The data has been collected from 5 different Southern Luzon regions, including Quezon, Rizal, Batangas, Laguna, and Cavite, from 23 rural regions. The Ensemble technique has been used in this research, including three machine learning algorithms, namely KNN, Naïve Bayes, and Regression Tree, which gave 97 % accuracy. In contrast, the stand-alone algorithm gave 90% accuracy [10]. The dataset used in this research was taken from Kaggle, where four machine learning algorithms, Logistic Regression, KNN, Random Forest, and ANN, achieved 60.51%, 60.98%, and 70.42%, 69.50% accuracy, respectively [11].

III. METHODOLOGY

The methodology used by the author in this study is depicted in Figure 1.

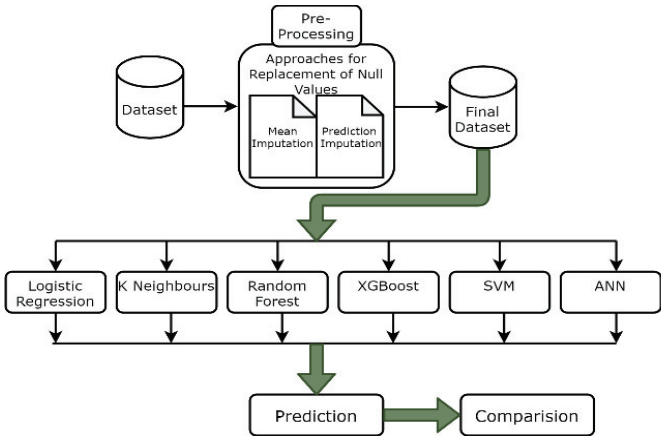

Fig. 1. Methodology Used

A. Dataset Information

Clean water is not only a necessity for human survival but also a fundamental human right and a cornerstone of any health protection strategy. Because of the stakes for health and development, this is an issue of national, regional, and even local significance. Furthermore, water and sewage systems investments are economically beneficial in some parts of the world. That's because the benefits of implementing the interventions, like fewer health problems and lower medical bills, far outweigh the costs. To conduct this experiment, the dataset is downloaded from Kaggle, which is open source and available for experiments [12]. There are 3276 unique bodies of water represented by quality metrics in the dataset. Dataset is downloaded in a comma-separated file format with the extension .CSV. Dataset attribute information with description is described in Table 1.

B. Tool and Hardware Details

In this study, the authors used Python language to execute the experiment. Jupyter notebook has been used to research and analyzes how different machine learning algorithms perform on the dataset. The investigation is conducted on the Ryzen 5 4600H processor, 8 GB of ram, and 512 GB of ROM. The dataset is downloaded in CSV format, so the dataset is ready for experiment through python libraries

C. Data Pre-processing

Values in real-world data are frequently redundant, and there is also a lot of noise. Before being fed into a model, the data must be cleaned up, and any missing values must be populated [18]. As part of the pre-processing phase, these issues are tackled to ensure reliable predictions. After the data has been cleansed, meaning that any erroneous information has been removed and any missing values have been populated, it must be transformed. The vast majority of the supervised learning algorithms are made to deal with nominal or cardinal data. Our present work applied a data transformation to the Kaggle dataset. By simplifying the dataset, we can transform it into a more digestible format, which improves the model's accuracy. However, the dataset contains some null values for some attributes, creating instability and difficulty correlating the data between attributes. To make the dataset stable for the experiment author proposes two approaches to replace the null values in the attribute column.

TABLE I. DATASET ATTRIBUTE INFORMATION

Attribute Name	Description
pH value	PH is responsible for checking the acid or base level of water. It also indicates the acidic or alkaline condition of the water. However, from 6.5 to 8.5 is the range of pH limit permissible by WHO.
Hardness	Calcium and magnesium salts are the main reason for the hardness of the water. Simply, it indicates the amount of these two salts present in water. Higher the hardness, the higher the number of salts. Hardness below 60 mg/l is considered soft water, the hardness between 60-120 is treated as moderately hard water, the hardness between 120-180 is considered hard water, and more than 180 mg/l is considered hard water.
Solids (Total dissolved solids - TDS)	This parameter is responsible for the color and taste of water by calculating the number of minerals present in water, like sulfate, chlorides, sodium, calcium, etc. Therefore, the higher the TDS(solids) in water, the higher the minerals dissolved in water, and the drinkable water has permissible TDS ranging from 500 mg/l to 1000 mg/L.
Chloramines	These are the disinfectants used to kill germs in the water. It forms by adding two substances, ammonia and chlorine. The drinkable water has a chlorine level of 4 mg/L.
Sulfate	Minerals, soil, and rocks are places where sulfates are generally found. Sulfate is heavily used in chemicals. More than 250 mg/L in water tastes bitter.
Conductivity	Conductivity in water refers to the number of substances dissolved in water. It is responsible for conducting electricity through water. Conductivity value should be less than 400 µS/cm, according to WHO.
Organic_carbon	This parameter refers to the amount of carbon in organic compounds in water, such as toluene, benzene, etc. TOC (Total Organic Compound) in drinking water should be < 2 mg/L.
Trihalomethanes	THMs are chemical products made during the treatment of water. Its amount is based on the amount of chlorine present in water for treatment. It should be less than 80 ppm in water to be treated as safe for drinking.
Turbidity	This parameter is responsible for the amount of solid matter in suspended form and water clarity. Turbidity should be less than 5 `NTU, according to WHO.
Portability	Indicates whether the quality of water is safe for drinking or not. 0 means not potable, and 1 tells potable.

Generally, some platforms' datasets have impurities like missing values, undesirable characters in rows, etc. The dataset used in this research has a large number of missing values. Usually, to handle missing values, we remove them, but if we have a large number of null values in our dataset, so we cannot remove them. Still, we can replace them with relevant entities, such as the measure of central tendencies having a mean, median, column mode, etc. Still, if our dataset has a higher number of null values, we should not just randomly fill them with these three statistical approaches. During the research, we filled the missing values in two approaches that affected our accuracy result by a significant margin.

D. Missing values in the used dataset

There is three column having null values in the dataset which have been used in this research, and the details about null values are:

Sulfate - 23.84 %

Ph - 14.99 %

Trihalomethanes - 4.95 %

E. Approach 1: Mean Imputation Approach

In the first approach, we have filled the missing values with the mean. There are three columns in the dataset we used with a large number of missing values in each column, so we replace the null values of each column with its mean of available values. The mean imputation approach process is depicted in Figure 2.

The mean approach worked as to suppose we have a feature(column) with six rows, and this feature has one missing value given as 14-20-13-16-12-NULL. Now the mean approach will work as follows:

Mean = Sum of not null rows(values) / count of not null rows(values)

Here,

$$14 + 20 + 13 + 16 + 12 / 5 = 75/5 = 15$$

Our pre-processed column will look like 14 20 13 16 12 15.

And we do not have missing values in this column. Still, the dataset in our research has many missing values after replacing null values with mean, which affected the relationship among the attributes. At the end of the model, the Author achieved the highest accuracy of 66.23% from ANN.

F. Approach 2: Predicting Imputation Approach

In the second approach, we have filled the missing values by prediction with the help of other attributes. Concretely, we first select the sulfate column's null values to fill with the rest columns of the dataset that do not have any null values. Then, sulfate is treated as Y on which prediction to implement, and the rest columns are treated as X, which are used to do prediction. Predicting imputation approach process is represented in Figure 3.

So, these X and Y columns were extracted from the parent data frame and stored in a new variable. Then, all the null values were removed, and now these columns have no null values. So, now, the sulfate column is treated as a training y variable, and the rest two are treated as training x variables and performed standardization on training x using a Standard scaler which brings the values of features in a particular range and is applied to Support Vector Machine to predict sulfate.

Grid search CV is the technique to get the best hyperparameter for a particular machine learning algorithm, giving the highest possible accuracy in SVM and producing the best hyperparameters. The model trained on this derived dataset and tested on the parent dataset with X parameters predicted the null values.

Now that we have the predicted values on behalf of null values that we stored in a variable, the sulfate parameter in the parent data frame will be concatenated by predicted sulfate values.

And this approach will apply to the rest column, which are pH and Trihalomethanes, which have null values, and extract other columns to predict pH and perform the second approach for the Trihalomethanes column. So at the end of this approach, we will get all the columns with no null values and 70.09% accuracy with ANN.

G. Machine Learning Algorithms

The author thoroughly reviewed the prior research before settling on the five classification algorithms listed below for use in the experiment.

- **Logistic Regression:** It is used for binary or one vs. many classifications' problems, which uses the sigmoid function as the activation function. The x-axis is for the data point, whereas the y-axis stands for the probability of the data point [13]. So, this algorithm determines the likelihood that the event will occur or not, and the possible lies in the range from 0 to 1. Hence it gives the outcome between this range. So it is a kind of regression algorithm for classification problems [14].

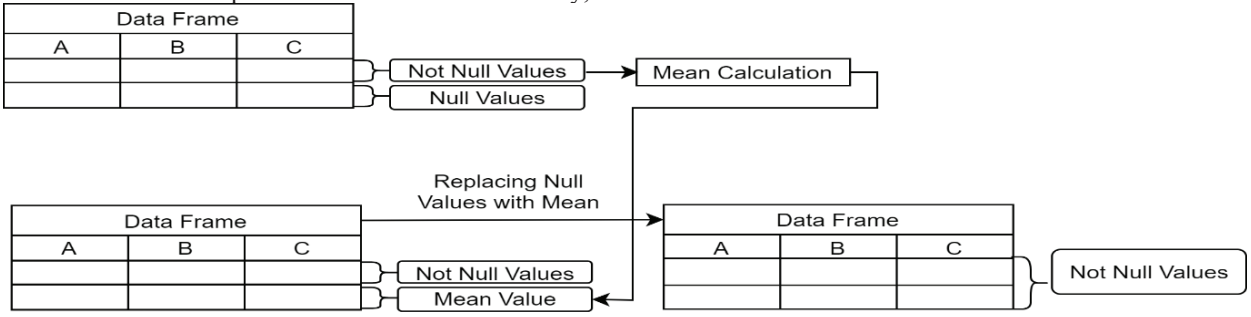

Fig. 2. Mean Imputation Approach

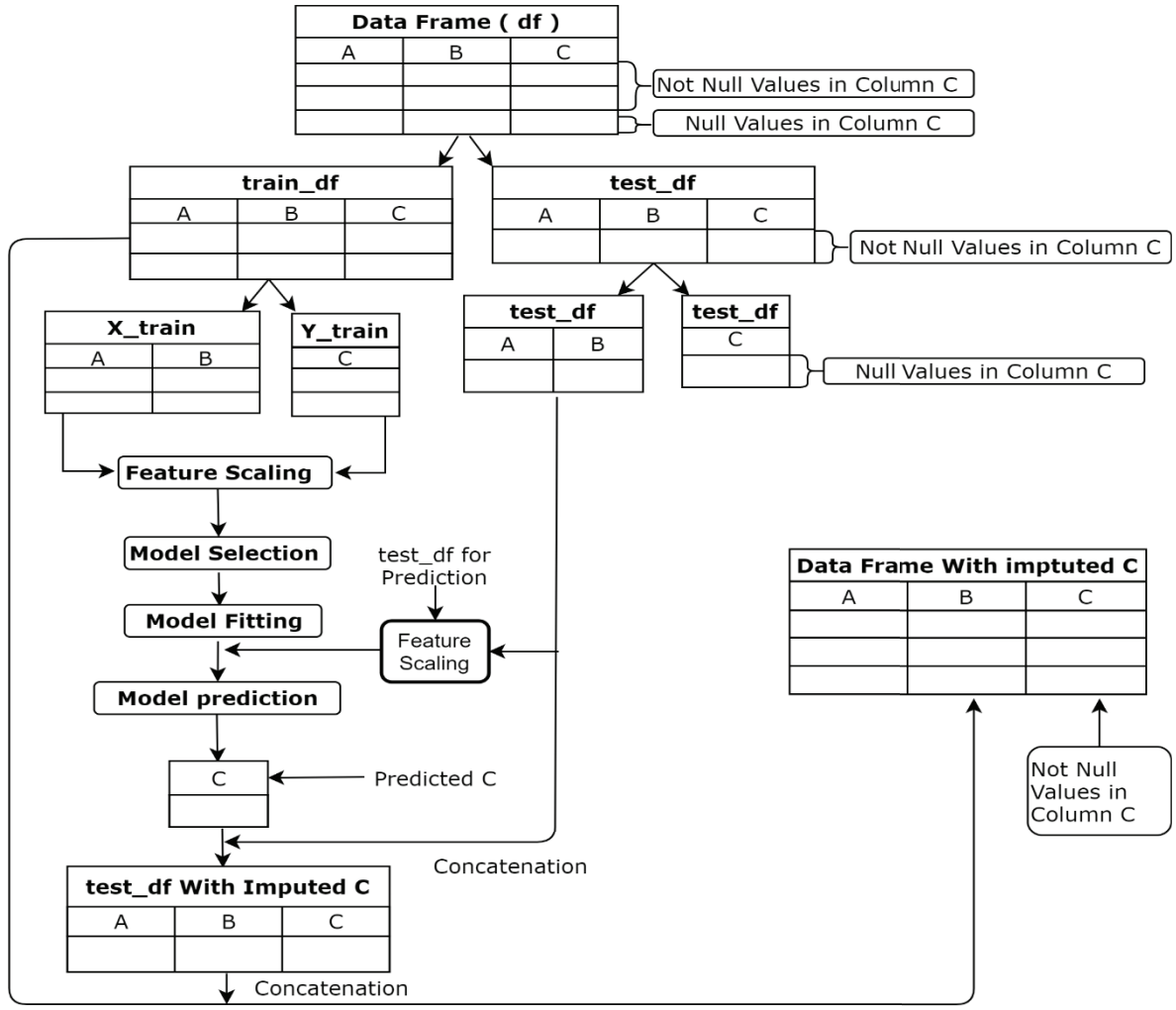

Fig. 3. Predicting Imputation Approach

- **Random Forest:** The combination of more than one decision tree can be used in classification or regression problems. It is the algorithm based on the ensemble technique. Each tree (Base tree) provided a sample of data from the dataset with sample replacement means some records may reassign to a different base learner called bootstrap. Each base learner predicts the classification class. Aggregation of outcome (mode in case of classification problem) is the final prediction [15].

- **Support Vector Machine:** This algorithm solves classification and regression problems. It creates a plane and two hyperplanes that are parallel to this plane to separate data of different categories [16]. Also, the hyperplane which has the highest margin will perform the best. Moreover, it can distinguish nonlinearly separable data using the SVM kernel, which can convert lower dimensions to higher dimensions to maintain a margin between the data of different classes [17].

- **K - Nearest Neighbor** is an unsupervised machine learning algorithm for regression and classification. In classification, if it has to find the class of a given datapoint, it will count the nearest neighbors to predict the class with the highest occurrence near it. In addition, this algorithm will consider the number of neighbors for analysis based on the user [18].

- **XG Boost** is based on a tree-based ensemble technique built on a gradient boosting framework. It is more complicated than other machine learning algorithms. However, it supports parallel computing in a single machine for learning in lesser time than different boosting algorithms like Gradient boost. Adaboost and Catboost. The algorithm gives better performance if its hyper-parameter is appropriately tuned. It worked on similarity scores for creating the tree. It is highly flexible, potable, and provides much more efficiency [19].

- **ANN:** This part of deep learning uses neurons to mimic the human brain for analyzing the pattern between the dataset's attributes. It creates a neural network with the data and interconnects with features for achieving the human brain capable of making decisions through previous information by examining how data is interconnected [20]. The neurons are created artificially for pattern analysis. That's why it is called an Artificial neural network consisting of multiple layers of neurons, where the layer which takes input from the user is known as

the input layer. This layer portrays the output result as the outer layer; the layers between these two are called hidden layers [21].

H. Evaluation Matrices

Predictions about whether or not the water is safe to drink are the focus of this study. The numbers 0 and 1 indicate whether or not the water is safe for consumption. Table 2 contains the confusion matrix.

According to the model's findings, TP can be used to produce potable water. The non-portability of water is denoted by the letter TN, and our model came up with the same conclusion. A false positive (FP) is a Type I error that occurs when a model predicts that the water is potable when it is not. False negatives are a type II error [22]. To put it another way, the model predicted that the water would be unusable, but in reality, the water is usable.

To determine the model's accuracy, the following formula must be used:

Accuracy = (TP+TN)/Total no. of instances (1)

Correctly predicted classes are counted as part of a recall's overall positive classes. The following is the formula:

Recall= (TP)/(TP+FN) (2)

The measure of precision is the number of actual positive classes compared to the total number of correctly predicted positive classes. The following is the formula that will be used to calculate the recall:

Precision = TP/(TP+FP) (3)

It becomes difficult to compare two models when one model's precision is low, and the other model's recall is high. If the inverse is true, the two parameters aren't considered when comparing models. When comparing models in these situations, the F-score is used. The F-score is calculated by taking the sum of the two values. Recall and precision can be measured simultaneously, which is extremely useful. The harmonic mean is frequently used instead of the arithmetic mean because it is not affected by outliers.

F-score= (2*Recall*Precision) / (Recall + Precision) (4)

TABLE II. CONFUSION MATRIX

Actual class\Predicted class	**C**	**Not in C**
C	**True Positives (TP)**	**False Negatives (FN)**
Not in C	**False Positives (FP)**	**True Negatives (TN)**

IV. EXPERIMENT AND RESULTS

The most critical and time-consuming step in extracting insights from data and improving the effectiveness of machine learning algorithms is pre-processing the data. To get the most information out of a dataset, we must deal with any contaminants that may exist. One impurity that may exist in a dataset is missing values, which has the potential to alter the meaning of the data. Therefore, we must know the business issue and the best solution to fill in the dataset's missing values.

No matter how sophisticated the machine learning algorithm is, it will produce undesirable results if the data is not adequately pre-processed. To get the best result, we discussed two methods in this research article for filling in missing or NULL values. In the first strategy, we used the mean to fill in the missing values, giving us a maximum accuracy of 66.23%, as depicted in Table 3.

TABLE III. MEAN IMPUTATION RESULT COMPARISON

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
Logistic Regression	51.76	61.45	53.29	57.88
K Neighbors	63.58	66.24	80.44	72.14
Random Forest	64.29	63.15	94.85	76.47
XGBoost	63.27	64.24	88.19	74.48
SVM	61.13	60.73	1.00	75.54
ANN	66.23	67.52	86.21	75.62

In the second strategy, we used prediction to fill in the missing values, leading to two algorithms scoring 70.09 % accuracy, as shown in Figure 4.

TABLE IV. PREDICTION IMPUTATION RESULT COMPARISON

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
Logistic Regression	61.03	61.96	1.00	76.25
K Neighbors	65.81	65.36	95.12	77.56
Random Forest	67.44	66.42	95.74	78.86
XGBoost	68.26	70.87	83.06	76.52
SVM	69.58	68.04	96.87	79.23
ANN	70.09	72.56	95.75	80.55

V. CONCLUSION

Water portability is essential for making water for drinking purposes, and various attributes are responsible for making water for drinking purposes. Null values in the dataset hinder the correlation of the data, so two approaches have been proposed, and there is an enhancement in accuracy through both methods. In the first approach, the mean Imputation author utilized the mean to fill in the missing values, which gave us the maximum accuracy of 66.23% using the ANN model. In the second approach, prediction imputation, the author used prediction to fill in the missing values, which led to the ANN algorithm scoring 70.09 % accuracy. Through this experiment, the author concludes that the prediction imputation approach is suitable for replacing the null values from a dataset. Still, the prediction imputation approach is a limitation: it takes much more execution time than the mean imputation approach but significantly increases performance. In the future, authors will implement this approach on various datasets and work on reducing the execution time of the experiment.

REFERENCES

[1] F. Ruiz-Garzón, M. del C. Olmos-Gómez, and L. I. Estrada-Vidal, "Perceptions of teachers in training on water issues and their relationship to the SDGs," *Sustainability*, vol. 13, no. 9, p. 5043, 2021.

[2] W. S. School, “The Water in You: Water and the Human Body,” 2022. https://www.usgs.gov/special-topics/water-science-school/science/water-you-water-and-human-body#:~:text=Water is of major importance,lungs are about 83%25 water (accessed Jul. 28, 2022).

[3] T. W. Bank, “World Water Day,” 2022. https://www.worldbank.org/en/country/india/brief/world-water-day-2022-how-india-is-addressing-its-water-needs (accessed Jul. 28, 2022).

[4] A. Kadiwal, “Water Portability Dataset.” 2014, Accessed: Jun. 28, 2021. [Online]. Available: https://www.kaggle.com/adityakadiwal/water-potability.

[5] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, “Water quality prediction using machine learning methods,” *Water Qual. Res. J.*, vol. 53, no. 1, pp. 3–13, 2018.

[6] O. K. Pal, “The Quality of Drinkable Water using Machine Learning Techniques,” *Int. J. Adv. Eng. Res. Sci.*, vol. 8, p. 5, 2021.

[7] USGS, “USGS Dataset,” 2021. https://www.usgs.gov/ (accessed Aug. 18, 2022).

[8] H. Lu and X. Ma, “Hybrid decision tree-based machine learning models for short-term water quality prediction,” *Chemosphere*, vol. 249, p. 126169, 2020.

[9] J. L. Lerios and M. V Villarica, “Pattern extraction of water quality prediction using machine learning algorithms of water reservoir,” *Int. J. Mech. Eng. Robot. Res.*, vol. 8, no. 6, pp. 992–997, 2019.

[10] M. I. Alipio, “Towards developing a classification model for water potability in Philippine rural areas,” *ASEAN Eng. J.*, vol. 10, no. 2, p. 24, 2020.

[11] D. Poudel, D. Shrestha, S. Bhattarai, and A. Ghimire, “Comparison of Machine Learning Algorithms in Statistically Imputed Water Potability Dataset.”

[12] S. Lee and D. Lee, “Improved Prediction of Harmful Algal Blooms in Four Major South Korea’s Rivers Using Deep Learning Models,” *Int. J. Environ. Res. Public Heal. 2018, Vol. 15, Page 1322*, vol. 15, no. 7, p. 1322, Jun. 2018, doi: 10.3390/IJERPH15071322.

[13] D. Kumar, V. Kukreja, V. Kadyan, and M. Mittal, “Detection of DoS attacks using machine learning techniques,” *Int. J. Veh. Auton. Syst.*, vol. 15, no. 3–4, pp. 256–270, 2020.

[14] Y. Wei and H. Hasan, “Application of Logical Regression Function Model in Credit Business of Commercial Banks,” *Appl. Math. Nonlinear Sci.*, 2021.

[15] C. Sharma, S. Sharma, M. Kumar, and A. Sodhi, “Early Stroke Prediction Using Machine Learning,” in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 890–894.

[16] D. Kumar and V. Kukreja, “Quantifying the Severity of Loose Smut in Wheat Using MRCNN,” in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 630–634.

[17] R. Zhao, “The Water Potability Prediction Based on Active Support Vector Machine and Artificial Neural Network,” in *2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR)*, 2021, pp. 110–114.

[18] P. Rozynek and M. Rozynek, “Water Potability Classification using Neural Networks,” 2021.

[19] F. Hadavimoghaddam, M. Ostadhassan, M. A. Sadri, T. Bondarenko, I. Chebyshev, and A. Semnani, “Prediction of water saturation from well log data by machine learning algorithms: boosting and super learner,” *J. Mar. Sci. Eng.*, vol. 9, no. 6, p. 666, 2021.

[20] V. Kukreja and D. Kumar, “Automatic Classification of Wheat Rust Diseases Using Deep Convolutional Neural Networks,” in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2021, pp. 1–6.

[21] S. Garg, A. Aleem, and M. M. Gore, “Employing Deep Neural Network for Early Prediction of Students’ Performance,” in *Intelligent Systems*, Springer, 2021, pp. 497–507.

[22] P. Das, S. Jain, C. Sharma, and S. Shambhu, “Prediction of Heart Disease Mortality Rate Using Data Mining,” 2021.