Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 24 | No. 4 | Article ID D1765 | 2025

Original Research Paper

https://doi.org/10.46488/NEPT.2025.v24i04.D1765

Open Access Journal

A Modification of the K-Nearest Neighbor Algorithm in the Assessment of Water Potability

Tanveer Ahmed Khan Fahim[1], Hasan Mahdi Mahi[2] and Adeeb Shahriar Zaman[1]†

[1]Department of Mathematics, University of Dhaka, Dhaka, Bangladesh
[2]Department of Mathematics and Natural Sciences, BRAC University, Dhaka, Bangladesh
†Corresponding author: Adeeb Shahriar Zaman; adeeb.math@gmail.com

Abbreviation: Nat. Env. & Poll. Technol.
Website: www.neptjournal.com

Received: 16-11-2024
Revised: 05-03-2025
Accepted: 10-03-2025

Key Words:

Water potability
Machine learning
K-nearest neighbors
Logistic regression
Random forest
Support vector machine
Artificial neural network

Citation for the Paper:

Fahim, T.A.K., Mahi, H.M. and Zaman, A.S., 2025. A modification of the K-nearest neighbor algorithm in the assessment of water potability. *Nature Environment and Pollution Technology*, 24(4), D1765. https://doi.org/10.46488/NEPT.2025.v24i04.D1765

Note: From 2025, the journal has adopted the use of Article IDs in citations instead of traditional consecutive page numbers. Each article is now given individual page ranges starting from page 1.

ABSTRACT

Water potability is crucial for public health, as access to clean and safe drinking water is vital for the prevention of waterborne diseases and promotion of overall well-being. Contaminated water poses significant health hazards, including gastrointestinal infections, chronic diseases, and potential outbreaks of life-threatening ailments, such as cholera. Dependable evaluation techniques are essential for detecting hazardous water sources and facilitating prompt action to reduce the hazards. In recent years, machine learning techniques have been versatile in solving classification problems, as they can analyze and discover hidden patterns in datasets that may be too complex for the human mind. In this study, we applied several machine learning techniques to predict the potability of a water body and attempted to modify one of these methods. The objective is to evaluate the models by testing their accuracy and propose a new model that is more advanced in terms of accurate prediction than the previous models. A dataset composed of nine features of a water body was used to examine the efficiency of the models in assessing water quality. By presenting a detailed comparison of the methods and results, we unlock a path for further modifications in the future, with the aim of further enhancing the performance and accuracy of the model.

INTRODUCTION

Covering more than 70% of the Earth's surface, water is the most crucial substance for all living organisms on Earth. We use it casually, while ignoring its essence. All living cells are composed of water, comprising approximately 25–85% of their composition (Kharat et al. 2017). Despite the crucial role of water in human lives, it is continuously being polluted. Water pollution is the contamination of water bodies. It can be polluted in various ways, although the main reason is human activity (Clark & Hakim 2013).

In 2022, a World Health Organization (WHO) study indicated that more than 2 billion people reside in places experiencing significant water stress, a situation expected to worsen in specific countries due to climate change and population expansion. Despite popular apprehension regarding emerging pollutants such as medicines, pesticides, polyfluoroalkyl compounds, and microplastics, it is important to emphasize that the most critical chemical hazards in drinking water continue to stem from substances such as arsenic, fluoride, and nitrate. Water pollution by microorganisms is a serious concern, as it results in the transmission of diseases such as cholera, dysentery, typhoid, and polio, which account for approximately 485,000 diarrheal-related fatalities each year (WHO 2022). According to the United Nations SDG 6 project report, by 2030, the health and livelihoods of 4.8 billion people may be jeopardized if the water quality and monitoring of aquatic systems are not improved. The global percentage of water bodies categorized as "good" decreased from 57 percent in 2017 to 56 percent in 2023 (UN Water 2024)

This has a higher fatality rate than incidents resulting from crimes, accidents, and acts of terrorism. Consequently, it is essential to offer novel methods for analyzing and, if feasible, predicting water quality. The water quality ecosystem has suffered due to rapid population growth, the industrial revolution, and the extensive application of pesticides and fertilizers (Cabral Pinto et al. 2019). Consequently, possessing models for forecasting water quality is beneficial for water monitoring.

Potable water is free from all toxins and hazardous microorganisms and is fit for consumption either directly by drinking or indirectly through food preparations. Despite being 70% covered with water, the Earth has a very limited source of potable water. The process of purifying water is complex and too costly for millions of people around the world, who live below the poverty line and do not have access to safe water.

Water can be purified in many different but effective ways (Agudelo-Vera et al. 2014). Other studies have been conducted using deep learning for water quality forecasting (Saeed et al. 2024), evaluating the performance of random forest, deep neural network, and long short-term memory for water quality management (Lee et al. 2022), detecting anomaly in drinking water (Dogo et al. 2019) and using a fuzzy logic model to evaluate water quality (Priya & Kumaravel 2024). Machine learning methods can help in this case by examining the different characteristics of water bodies and predicting whether the water is fit for human consumption and other uses.

The idea behind Machine learning algorithms can learn from a given dataset by finding hidden patterns and making decisions without human intervention. Logistic regression helps classify tasks by estimating the probability of a binary outcome after inspecting the regressors or features. In contrast, the k-nearest neighbors, which is a non-parametric method, labels a point based on the closest neighbors of that specific point. Both algorithms were used in this study to assess the potability of water. Both of these algorithms, along with support vector machines, random forests, and artificial neural networks, were used in our study to assess the potability of water. Many research studies have been carried out in recent times in detecting water potability by machine learning models (Kaddoura 2022, Patel et al. 2022, Dalal et al. 2022), and many are still in progress. One of these studies, conducted by Poudel et al. (2022), compared four machine learning algorithms for a statistically imputed real-world water potability dataset. The study was performed using 20% of the data as test data in a randomized order, and checking the accuracy of each method for this test dataset. Logistic Regression, KNN, Random Forest and Artificial Neural Network obtained accuracies of 60.51%, 60.98%, 70.42% and 69.50% respectively (Poudel et al. 2022). Our goal is to provide insights into the algorithms while mentioning some of their robustness and drawbacks, and, if possible, modifying a model to obtain better accuracy than that obtained in the study mentioned above. By performing a comparative analysis of these models, we may help future research studies enhance the performance of a model to accurately predict potability.

The following chapters provide an outline of the necessary terminologies, brief details about the dataset, the findings of our study, and a short discussion of them.

MATERIALS AND METHODS

Preliminaries

Machine learning: Machine learning, a subset of AI, is defined as the field of study in which a computer can learn hidden patterns from given data without being explicitly programmed. It analyzes the structures in data to continue learning, reasoning, and decision-making without human intervention.

Consider a fake news detection program. The machine was given samples of both real and fake news. These data were considered as training data. The machine learns from the samples by carefully observing various features (including n-grams, punctuation, grammar, and readability). The objective is to detect whether the latest news is real or fake using prior experience. Its performance can be evaluated by checking how many news items are labeled correctly (Alghamdi et al. 2024).

There are various categories of machine learning algorithms. However, in this study, only supervised learning and its classification are discussed.

Supervised and unsupervised learning: Two primary branches of machine learning are supervised and unsupervised learning. In supervised learning, labeled data are used to train the algorithm to make accurate predictions. Data that has been assigned a label or class is known as labeled data. For instance, if a dataset image of dogs and cats has the labels "dogs" and "cats," it would be classified as a labeled dataset.

Unsupervised learning involves machine learning from unlabeled data. The machine analyzes the data to discover hidden patterns and performs data clustering on unlabeled data. In the previous example, if the tags "dogs" and "cats" were not given, the machine would have analyzed the features of those animals and labeled them in different classes based on the similarities they possess. This mirrors

the learning process of humans, where something is not known beforehand (GeeksforGeeks 2024a).

We will not discuss unsupervised learning further in this paper. Regression and classification are the two major aspects of supervised learning.

- Classification involves dividing the dataset into distinct classes based on different parameters. Algorithms such as K-nearest neighbors and decision trees fall under the classification aspect.
- Regression is involved when the data show a strong relationship between the independent and dependent variables. Patterns are identified within the training dataset. Linear regression, polynomial regression and logistic regression are some of the popular algorithms (Shalev-Shwartz 2014).

We will now discuss the K-nearest neighbors (KNN), which is a classification algorithm, and logistic regression, which is a regression algorithm. Later, we will compare these two algorithms for a specific problem and attempt to modify any algorithm, if possible, to obtain a better result.

K-nearest-neighbors (KNN): The idea behind this method is that points with the same characteristics will be closer in terms of distance, and thus their output will be the same. This method can be implemented for binary classification problems.

Metric

A metric (or distance function) on a set X is a real-valued function

$$d: X \times X \to \mathbb{R}$$

that satisfies the following four properties for all $x, y, z \in X$

1. $d(x,y) \geq 0$
2. $d(x,y) = 0 \leftrightarrow x = y$
3. $d(y,x) = d(x,y)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

Some popular distance metrics are as follows:

- **Euclidean Distance:** $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- **Manhattan Distance:** $d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$
- **Chebyshev Distance:** $d(x,y) = \max_{i=1,2,\ldots,n}|x_i - y_i|$

In the KNN method, we use one of the distant metrics to measure the distance between the test and training data.

First, the data were considered without a label. Then, it is assigned a label based on the label of points closer to it. This is done because it is assumed that the data have a higher chance of being in the same group as the points closest to them.

It is crucial to know which k-value to select. Generally, this depends on the given dataset. k = 1 will work perfectly if the dataset contains a strong pattern. However, in the real world, most datasets contain ambiguity. In these cases, k = 1 no longer works. Generally, odd values of k are suitable to avoid situations where ties occur between two groups. Cross-validation methods help select the best k value depending on the input data (GeeksforGeeks, 2024b).

Modification of the KNN method: A modification of the existing KNN method is possible by changing the distance metric. Instead of using the Euclidean distance formula, we propose the following formula:

$$d(x,y) = \frac{\sqrt{\sum_{i=1}^{n} k_i(x_i - y_i)^2}}{\sqrt{\sum_{i=1}^{n} k_i^2}}$$

Initially a vector $\vec{k}$ of the coefficients are chosen. Then, we apply the gradient descent method to find the optimal k-coefficients and ultimately use these optimal values to predict the outcome.

Here, we can easily prove that $d(x, y)$ is a metric and thus suitable for using as a distance function. The proof is as follows.

1. The numerator $\sqrt{\sum_{i=1}^{n} k_i(x_i - y_i)^2}$ is the square root of a sum of squared terms with k_i's being positive real numbers. The denominator $\sqrt{\sum_{i=1}^{n} k_i{}^2}$ is a constant and always positive if at least one $k_i \neq 0$.. Therefore, both the numerator and denominator terms are non-negative. So, $d(x,y) \geq 0$ for all x, y.
2. $d(x,y) = 0 \leftrightarrow k_i(x_i - y_i)^2 = 0 \leftrightarrow x_i = y_i$ for all i.
3. $d(y,x) = \frac{\sqrt{\sum_{i=1}^{n} k_i(y_i - x_i)^2}}{\sqrt{\sum_{i=1}^{n} k_i^2}} = \frac{\sqrt{\sum_{i=1}^{n} k_i(x_i - y_i)^2}}{\sqrt{\sum_{i=1}^{n} k_i^2}} = d(x,y)$
4. From the Minkowski inequality with p = 2, it follows.

$$\sqrt{\sum_{i=1}^{n} c_i(a_i + b_i)^2} \leq \sqrt{\sum_{i=1}^{n} c_i a_i{}^2} + \sqrt{\sum_{i=1}^{n} c_i b_i{}^2}$$

Replacing c_i by k_i and putting $a_i = x_i - y_i$ and $b_i = y_i - z_i$, we get

$$\sqrt{\sum_{i=1}^{n} k_i(x_i - z_i)^2} \leq \sqrt{\sum_{i=1}^{n} k_i(x_i - y_i)^2} + \sqrt{\sum_{i=1}^{n} k_i(y_i - z_i)^2}$$

Dividing both sides by $\sqrt{\sum_{i=1}^{n} k_i{}^2}$ we get,

$$d(x,z) \leq d(x,y) + d(y,z)$$

So, $d(x, y)$ is a metric.

Logistic regression: In logistic regression, we used the sigmoid function shown in Fig. 1. It is defined by,

$$h(z) = \frac{1}{1+e^{-z}}$$

This method can also be used for classification tasks. It produces the probability of making a binary choice. The target variable was either 0 or 1.

The idea behind this model is based on a linear regression. The equation of linear regression is given by,

$$Y = \theta^T x \quad \ldots(2.1)$$

The output y is used as the argument of the sigmoid function to get the estimated probability,

$$\hat{P} = f_\theta(x) = h(\theta^T x) \quad \ldots(2.2)$$

In the linear regression model, the Mean Squared Error (MSE) cost function is,

$$MSE(x, f_\theta) = \frac{1}{m_i}\sum_{i=1}^{n}\left(\theta^T x^{(i)} - y^{(i)}\right)^2 \quad \ldots(2.3)$$

For logistic regression, the term $\theta^T x^i$ is replaced by the functional value of it, i.e. $h(\theta^T x^i)$ or simply $f_\theta(x^i)$. So, the error function is,

$$MSE(x, f_\theta) = \frac{1}{m_i}\sum_{i=1}^{n}\left(f_\theta(x^{(i)}) - y^{(i)}\right)^2 \quad \ldots(2.4)$$

The primary objective is to obtain the vector θ^T Minimizing the error function in (2.4). Now, we check if $\hat{P} > 0.5$. We define,

$$\hat{y} = \begin{cases} 1 & if \quad \hat{P} > 0.5 \\ 0 & if \quad \hat{P} < 0.5 \end{cases}$$

Logistic regression is applicable in scenarios such as predicting the win percentage of a team in a football match, the likelihood of rain on a specific date, and the diagnosis of diseases.

Support vector machine: Support Vector Machine (SVM) is a supervised learning algorithm that is effective for solving binary classification problems owing to its strength in high-dimensional spaces. It works by finding the optimal hyperplane that best separates the data points from different classes.

A hyperplane is a decision boundary that separates different classes in feature space. SVM aims to maximize the margin, which is the distance between the hyperplane and the closest data points of each class. These points, which are closest to the hyperplane, are called support vectors, and they define the optimal hyperplane.

SVM uses a kernel function to transform the data into a higher-dimensional space, where it separates the space into two regions for binary problems. Some common kernel functions are as follows:

- **Linear Kernel:** Used when data is linearly separable.
- **Polynomial Kernel:** Data are mapped into polynomial space.

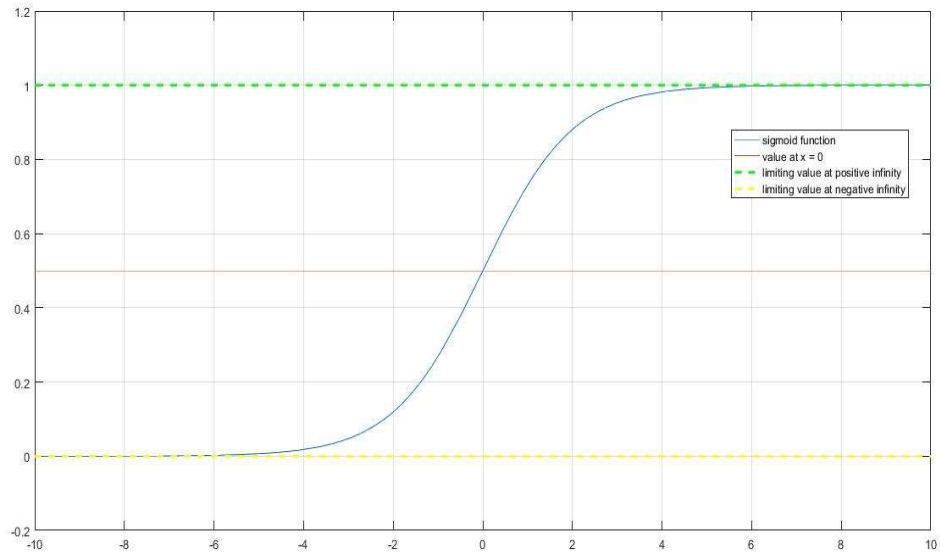

Fig. 1: Sigmoid Function.

- **Radial Basis Function (RBF):** Works well for complex, nonlinear data.

SVM uses a hinge loss function designed to maximize the margin while penalizing misclassified points. The loss function is given by

$$L = \sum_{i=1}^{n} \max(0, 1 - y_i (w * x_i + b)) \quad \#(1)$$

- x_i = training sample
- y_i =class label (+1 or -1)
- w and b = hyperplane parameters

The hinge loss ensures zero loss when the SVM correctly classifies the data points, and for each misclassified point, a penalty is added.

Random forest: Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It is built on decision trees and improves the accuracy while reducing overfitting.

Random Forest works by creating a " Forest " of multiple decision trees and aggregating their predictions. The algorithm randomly selects samples from the dataset to train every tree. Each tree learns from a different subset of data to reduce the variance and overfitting. These decision trees are built using a subset of features, and they split the data based on the most informative feature at each step. Their predictions were combined for the final output.

Random Forest does not use a single loss function, as in the case of SVM. Instead, it minimizes errors by averaging the predictions from multiple trees.

Artificial neural network (ANN): An Artificial Neural Network (ANN) is a machine learning model inspired by the structure of the human brain. Consisting of many layers of interconnected nodes (neurons) that process input data and learn patterns through training, ANNs are widely used in fields such as image recognition, Natural Language Processing (NLP), medical diagnosis, and financial forecasting.

ANN consists of three main types of layers.

- **Input Layer:** Receives raw data and assigns a weight to each input before passing it to the next layer.
- **Output Layer:** Produces the final prediction and uses an activation function. ReLU (Rectified Linear Unit (ReLU), sigmoid, and hyperbolic tangent (tanh).
- **Hidden layers:** Computations are performed by applying weights, biases, and activation functions. A Deep Neural Network is formed by multiple hidden layers. Each neuron computes:

$$z = W * X + B$$

Where,

W = weight matrix

X = input

B = bias

Table 1: provides the computational efficiency (time complexity) of different machine learning methods (Gupta 2023, Reddy 2023).

Where,

- n = Number of training samples
- d = Number of features
- s = Number of support vectors
- m = Number of trees
- k = Depth of tree
- i = Number of iterations
- h = Number of hidden units
- p = d*h

Attributes

The database we used contains information on 3276 water bodies, including lakes, rivers, and oceans (Kaggle 2024). It contains nine features of the 700 water bodies. Those are:

pH: pH level of the water body

Hardness: Dissolved calcium and magnesium salts

Solids: Dissolved materials, including both organic and inorganic

Chloramines: Disinfectant created by mixing chlorine with ammonia

Sulfate: Ion produced from sulfide minerals

Conductivity: Ability to conduct electricity

Organic Carbon: Carbon from living organisms

Trihalomethanes: By-product of water treatment

Turbidity: A Measure of Water Clarity

Potability: 1 indicates safe water, 0 indicates unsafe water

Table 1: Time complexity of different methods.

Method	Training Time	Prediction Time
K-Nearest Neighbor	$O(1)$	$O(n * d)$
Logistic Regression	$O(n * d)$	$O(d)$
Support Vector Machine	$O(n^2)$	$O(s * d)$
Random Forest	$O(m * n * logn * d)$	$O(m * k)$
Artificial Neural Network	$O(i * n * p)$	$O(p)$

There are 491 missing pH value data, which is approximately 15%, 781 missing sulfate data, which is approximately 24%, and 5% trihalomethane data are missing. We removed all rows of data points in which at least one feature value was missing. After removing all the missing value rows, we ended up with 2011 data points.

Because the dataset was too large, we randomly selected some data points to ensure that the models performed efficiently. To ensure a random selection of data points, we applied the RAND function to an additional column in Excel for the 3,276 water body data points. Because these values were generated randomly, we sorted the dataset in ascending order based on the assigned random numbers. Finally, we selected the top 700 data points for this study. We repeated this process multiple times to train the model on different datasets and found more or less the same accuracy each time.

RESULTS AND DISCUSSION

The data was partitioned into two classes. The training data comprised 90% of the data, and the remaining 10% was labeled as test data. The logistic regression model successfully predicted the potability 45 times out of 70, which resulted in an accuracy of 64.29%. We used a polynomial as the kernel function in SVM and tanh (hyperbolic tangent) as the activation function in ANN. We used 500 hidden layers for the ANN model and 1000 decision trees for the Random Forest, where we trained each tree on the entire training dataset. The accuracies of the Support Vector Machine (SVM), random forest, and Artificial Neural Network (ANN) are 67.14%, 64.29%, and 67.14%, respectively.

In case of the KNN model in Fig. 2, k = 1 and k = 3 led to the accuracies of 67.14% and 58.57% respectively. The accuracy then increased gradually and reached a peak accuracy of 71.43% for k = 5. The accuracy decreased and remained at approximately 62% in the long run. Primarily, the high accuracy for smaller k values was due to overfitting.

The accuracy slightly increased for k = 4 and k = 5, and then began to decrease, while remaining above 60%. (95% Confidence Interval for Accuracy: (55.5%, 62.8%))

The modified KNN model in Fig. 3 gave us an accuracy of 60% for k = 1 and 64.29% for k = 3. At first, it appears that this modified algorithm does not improve the accuracy at all. However, if we check the accuracy for the value of k up to 30, we see that the performance is greatly improved.

For most k-values, the accuracy was above 67%, which is better than that of the other models (95% Confidence Interval for Accuracy: 62.8%, 70.3%).

This new model achieved a 70% accurate result at k = 9. For the majority of the k-values, the accuracy remains above 67% and the peak accuracy is gained at k = 4, which is 77.14%. Subsequently, the accuracy slightly decreased, but as the value of k continued to increase, the accuracy reached a steady level of approximately 67%.

Table 2: presents the accuracy of the methods discussed throughout this paper. We can clearly observe that our modification of the KNN model is better at predicting water potability than all the other models.

We also used the 10-fold cross-validation technique to check whether the improved accuracy of the modified

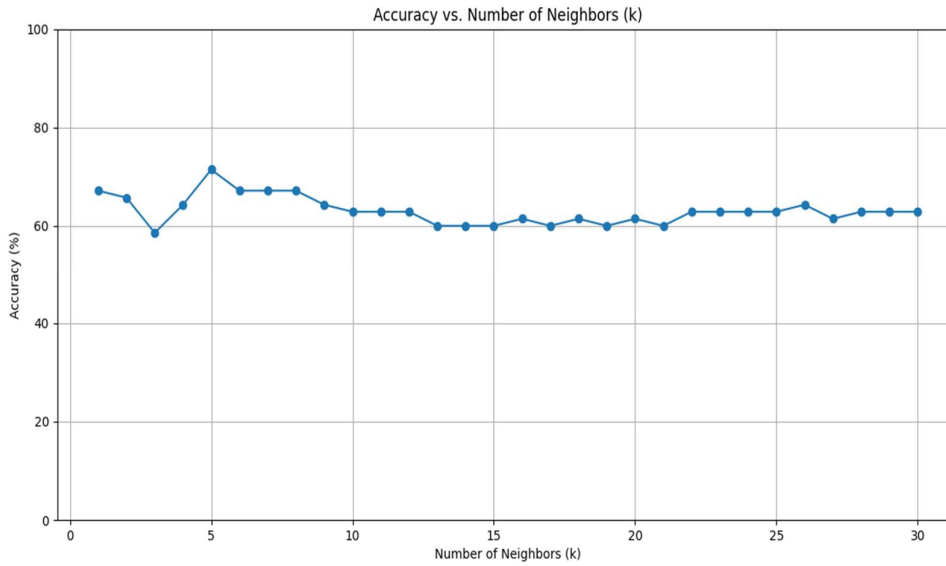

Fig. 2: Graph of accuracy in the KNN method for k-value of up to 30.

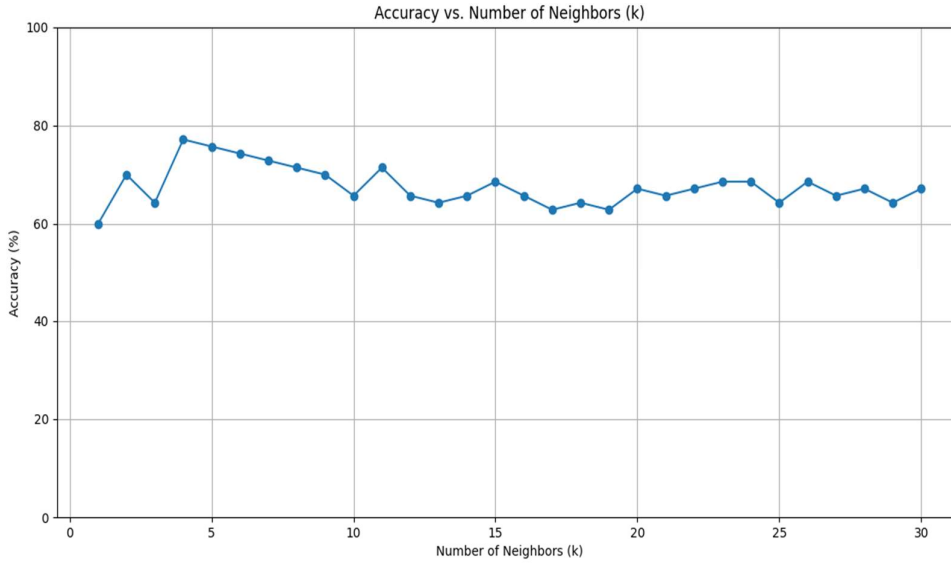

Fig. 3: Graph of accuracy in the Modified KNN method for k-value up to 30.

Table 2: Accuracies of different machine learning methods.

Method	Accuracy [%]
Logistic Regression (LR)	64.29%
K-Nearest Neighbor (KNN)	62.86% (K = 30)
Support Vector Machine (SVM)	67.14%
Random Forest (RF)	64.29%
Artificial Neural Network (ANN)	67.14%
Modified KNN	70.00% (K = 9)

model was significantly better than that of the other models. In addition, there is a slight chance of bias or overfitting if the model is trained on a single training dataset. To tackle this problem, we divided the entire dataset into 10 different folds and used one of the folds as the test data and the remaining nine as the training data. This reduces the chance of overfitting and provides a much more stable accuracy result. To do this, we compared the 10 accuracies of the modified model with those of the other models by calculating the t-statistic and p-value. If the t-statistic value is positive, our modified model is better than the other model, and if the value is negative, our modified model is worse. If the p-value was less than 0.05, the difference in accuracy was significant. The mean accuracy of the modified model was 66.57%, and the confidence interval was (62.8%, 70.3%). Table 3 shows the mean accuracy of the 10 folds of each model (except the modified model) along with the t-statistic and p-value, the confidence interval (95%), and a comment on whether the modified model is significantly better or not:

We attempted to remove all possible randomness associated with each model and trained the model on the entire training dataset. However, there is a slight inherent randomness in the ANN model; therefore, we checked eight different random states. Our modified model was better than all of those models, and in six of the cases, the accuracy was significantly better. Therefore, we can confirm that the modification of the model is successful, as the model is significantly better than the other models. Further changes in this model can be made to improve accuracy by using different distance metrics.

CONCLUSIONS

We attempted to modify the existing KNN method by changing the distance metric and applying the idea of weighted distance to the model. Our purpose was to build

Table 3: T-statistic and P-value of different methods.

Method	Mean Accuracy	t-statistic	p-value	Comment	CI
KNN	59.14%	6.0900	0.0002	Significantly better	(55.5%, 62.8%)
LR	61.43%	3.2208	0.0105	Significantly better	(58.1%, 64.8%)
SVM	62.86%	2.8216	0.0200	Significantly better	(60.5%, 65.2%)
RF	60.86%	3.7201	0.0048	Significantly better	(57.8%, 63.9%)
ANN	63%	2.3886	0.0407	Significantly better	(59.0%, 67.0%)

a model that can yield much better accuracy in predicting the purity of water than the methods mentioned above. The modified KNN model was significantly better than the KNN model in predicting water potability, as there was a significant increase in accuracy. The modified KNN model also offered better accuracy than the Logistic Regression, SVM, Random Forest and ANN models. Considering both accuracy and the fact that the modified KNN model is easier to work with and implement than Random Forest and ANN, we can declare that our modified model is an upgrade from those models. However, there is still room for improvement in this area. As the future of machine learning looks promising, further research in this field will help us build a model with higher accuracy and precision. This study will aid in that journey by providing insights into the algorithms for detecting water potability.

REFERENCES

Agudelo-Vera, C., Blokker, M., Vreeburg, J., Bongard, T., Hillegers, S. and Van Der Hoek, J.P., 2014. Robustness of the drinking water distribution network under changing future demand. *Procedia Engineering*, 89, pp.339-346.

Alghamdi, J., Luo, S. and Lin, Y., 2024. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications*, 83(17), pp.51009-51067.

Cabral Pinto, M.M., Ordens, C.M., Condesso de Melo, M.T., Inácio, M., Almeida, A., Pinto, E. and Ferreira da Silva, E.A., 2020. An inter-disciplinary approach to evaluate human health risks due to long-term exposure to contaminated groundwater near a chemical complex. *Exposure and Health*, 12, pp.199-214.

Clark, R.M. and Hakim, S., 2014. Securing Water and Wastewater Systems. Cham: Springer International, pp.978-3.

Dalal, S., Onyema, E.M., Romero, C.A.T., Ndufeiya-Kumasi, L.C., Maryann, D.C., Nnedimkpa, A.J. and Bhatia, T.K., 2022. Machine learning-based forecasting of potability of drinking water through adaptive boosting model. *Open Chemistry*, 20(1), pp.816-828.

Dogo, E.M., Nwulu, N.I., Twala, B. and Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water Journal*, 16(3), pp.235-248.

GeeksforGeeks, 2024a. Supervised Learning. Retrieved November 14, 2024, from https://www.geeksforgeeks.org/supervised-unsupervised-learning/

GeeksforGeeks, 2024b. K Nearest Neighbor. Retrieved November 14, 2024, from https://www.geeksforgeeks.org/k-nearest-neighbours/

Gupta, R., 2023. Time Complexity in Machine Learning. Retrieved March 3, 2025, from https://medium.com/@riteshgupta.ai/time-complexity-in-machine-learning-4c253919e871

Kaddoura, S., 2022. Evaluation of machine learning algorithms on drinking water quality for better sustainability. *Sustainability*, 14(18), p.11478.

Kaggle, 2024. Dataset of Water Potability. Retrieved November 14, 2024, from https://www.kaggle.com/datasets/adityakadiwal/water-potability

Kharat, M., Du, Z., Zhang, G. and McClements, D.J., 2017. Physical and chemical stability of curcumin in aqueous solutions and emulsions: Impact of pH, temperature, and molecular environment. *Journal of Agricultural and Food Chemistry*, 65(8), pp.1525-1532.

Lee, H.W., Kim, M., Son, H.W., Min, B. and Choi, J.H., 2022. Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea. *Journal of Hydrology: Regional Studies*, 41, p.101069.

Patel, J., Amipara, C., Ahanger, T.A., Ladhva, K., Gupta, R.K., Alsaab, H.O., Althobaiti, Y.S. and Ratna, R., 2022. A machine learning-based water potability prediction model by using the synthetic minority oversampling technique and explainable AI. *Computational Intelligence and Neuroscience*, 2022(1), p.9283293.

Poudel, D., Shrestha, D., Bhattarai, S. and Ghimire, A., 2022. Comparison of machine learning algorithms in statistically imputed water potability dataset. *Journal of Innovations in Engineering Education*, 5(1), pp.38-46.

Priya, M. and Kumaravel, R., 2024. Fuzzy logic harmony in water: Mamdani inference system applied to evaluate pristine pond water quality. *Nature Environment and Pollution Technology*, 23(3), pp.1775-1782.

Reddy, A., 2023. Machine Learning. Retrieved March 3, 2025, from https://anudeepareddy-s.medium.com/machine-learning-c85118ec9a1b

Saeed, A., Alsini, A. and Amin, D., 2024. Water quality multivariate forecasting using deep learning in a West Australian estuary. *Environmental Modelling & Software*, 171, p.105884.

Shalev-Shwartz, S. and Ben-David, S., 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press.

UN Water, 2024. Progress on Ambient Water Quality – 2024. Retrieved November 14, 2024, from https://www.unwater.org/publications/progress-ambient-water-quality-2024-update

World Health Organization (WHO), 2022. Drinking Water. Retrieved November 11, 2024, from https://www.who.int/news-room/fact-sheets/detail/drinking-water