Prepublished copy

Nature Environment & Pollution Technology

This is a peer-reviewed prepublished version of the paper

Original Research

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

Key Words	Water Potability, Machine Learning, K-Nearest Neighbors, Logistic Regression, Modified K-Nearest Neighbors, Random Forest, Support Vector Machine, Artificial Neural Network.
DOI	https://doi.org/10.46488/NEPT.2025.v24i04.D1765 (DOI will be active only after the final publication of the paper)
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), p. D1765 https://doi.org/10.46488/NEPT.2025.v24i04.D1765

ABSTRACT

Water potability is a crucial necessity for public health, as access to clean and safe drinking water is vital for the prevention of waterborne diseases and the promotion of overall well-being. Contaminated water poses significant health hazards, including gastrointestinal infections, chronic diseases, and potential outbreaks of life-threatening ailments like cholera. Dependable evaluation techniques are essential for detecting hazardous water sources and facilitating prompt action to reduce hazards**.** In recent years, machine learning techniques have been versatile in solving classification problems as they can analyze and discover hidden patterns in the datasets which can possibly be too complex for human minds. In this study, we applied several machine learning techniques for predicting the potability of a water body and attempted a modification of one of those methods. The objective is to evaluate the models by testing their accuracies and propose a new model which is more advanced at predicting accurately than the previous models. A dataset composed of 9 features of a water body is used to examine the efficiency of the models in assessing water quality. By presenting a detailed comparison of the methods and the results, we unlock a path for more modification 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. We use it casually while ignoring its essence. All living cells are made up of water around 25 - 85% (Kharat et al. 2017). In spite of water playing such a crucial role 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 activities (Clark et al. 2013).

In 2022, World Health Organisation (WHO) study indicates 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 micro plastics, it is important to emphasize that the most critical chemical hazards in drinking water continue to stem from substances like 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 United Nations SDG 6 project report, by 2030, the health and livelihoods of 4.8 billion humans may be jeopardised if water quality and monitoring of aquatic systems are not improved. The global percentage of water bodies categorised as "good" decreased from 57 percent in 2017 to 56 percent in 2023. (UN Water,2024)

This has a greater 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 the swift 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 quite beneficial for water monitoring.

Potable water means water being free from all toxins and hazardous microorganisms and fit to be consumed either directly by drinking or indirectly by food preparations. In spite of being 70% covered with water, the earth has very limited source of potable water. The process of purifying the water is complex and too costly for millions of people around the world 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). Some other studies have been conducted by utilizing deep learning for water quality forecasting (Amina et al. 2024), evaluating the performance of random forest, deep neural network, long short-term memory for the water quality management (Hye et al. 2022) and using fuzzy logic model to evaluate water quality (Priya and 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 usages or not.

The idea behind machine learning algorithms is that it can learn from a given dataset by finding the hidden patterns and make decisions without human intervention. Logistic regression helps in classifying tasks by estimating probability of a binary outcome after inspecting the regressors or features. On the contrary, k-nearest neighbors which is a non parametric method label a point based on the closest neighbors of that specific point. Both of these algorithms have been used in our study to assess the potability of water. Both of these algorithms

along with Support Vector Machine, Random Forest and Artificial Neural Network have been 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, S, 2022, Patel et al. 2022, Dalal et al. 2022) and many are still in progress. One of those studies, done by Poudel et al. in 2022, compares four machine learning algorithms for a statistically imputed real-world water potability dataset. The study was done by 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 D., Shrestha D. et al. 2022). Our goal is to provide insights about the algorithms while mentioning some of their robustness and drawbacks and, if possible, modifying a model to get better accuracy than those obtained in the study mentioned before. By performing a comparative analysis on these models, we might help the future research studies on enhancing the performance of a model to accurately predict the 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.

2. PRELIMINARIES

2.1 Machine Learning

A subset of AI, machine learning is defined as the field of study where a computer can learn hidden patterns from given data without being explicitly programmed. It analyzes the structures in data to keep on learning, reasoning and decision-making without the help of human intervention.

Consider a fake news detection program. The machine is given samples of both real and fake news. These data are considered 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 a latest news is real or fake using the prior experience. Its performance can be evaluated by checking how many news is labeled correctly (J. Alghamdi et al. 2024).

Variety of categories exist for machine learning algorithms. However, in this paper, only supervised learning and its classification will be discussed.

2.2 Supervised and Unsupervised Learning

Two of the primary branches in the field of machine learning are supervised and unsupervised learning. In supervised learning, labeled data is used to train the algorithm to make accurate prediction. Data which have been assigned a label or a 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 (Supervised Learning 2024).

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

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

- Classification involves dividing the datset into distinct classes based on different parameters. Algorithms like K-nearest neighbors, Decision trees fall under the classification aspect.

- Regression is involved when data shows a strong relationship between the independent and dependent variables. Pattern is identified within the training dataset. Linear regression, polynomial regression and logistic regression are some of the popular algorithms (Shavel-Shawrtz S. 2014).

Now, we will discuss about K-nearest neighbors (KNN) which is a classification algorithm and logistic regression which is a regression algorithm. Later, we will make a comparison between these two algorithms for a specific problem and try to modify any algorithm, if possible, to obtain a better result.

2.2.1 K-Nearest-Neighbors (KNN)

The idea behind this method is that points with the same characteristic will be closer in terms of distance and thus their output will be same also. This method can be implemented in 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 of the popular distance metrics are:

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

At first, the data is considered without a label. Then, it is assigned a label based on the label of the points closer to it. This is done because it is assumed that the data has a higher chance of being in the same group with the points it is closest to.

It is crucial to know which k-value to choose. Generally, it depends on the given dataset. k = 1 will work perfectly fine if the dataset contains strong pattern. But in the real world, most of the dataset contain ambiguity. In those cases, k = 1 no longer work. Generally, the odd values of k are suitable in order to escape situations where tie occurs between two groups. Cross validation methods help for selecting the best k value depending on the input data (K Nearest Neighbor 2024).

2.2.2 Modification of KNN Method

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

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

Initially a vector $\vec{k}$ of the coefficients are chosen. Then we apply 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 $k_i \neq 0$. So, 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 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.

2.2.3 Logistic Regression

In logistic regression, we use sigmoid function. It is defined by,

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

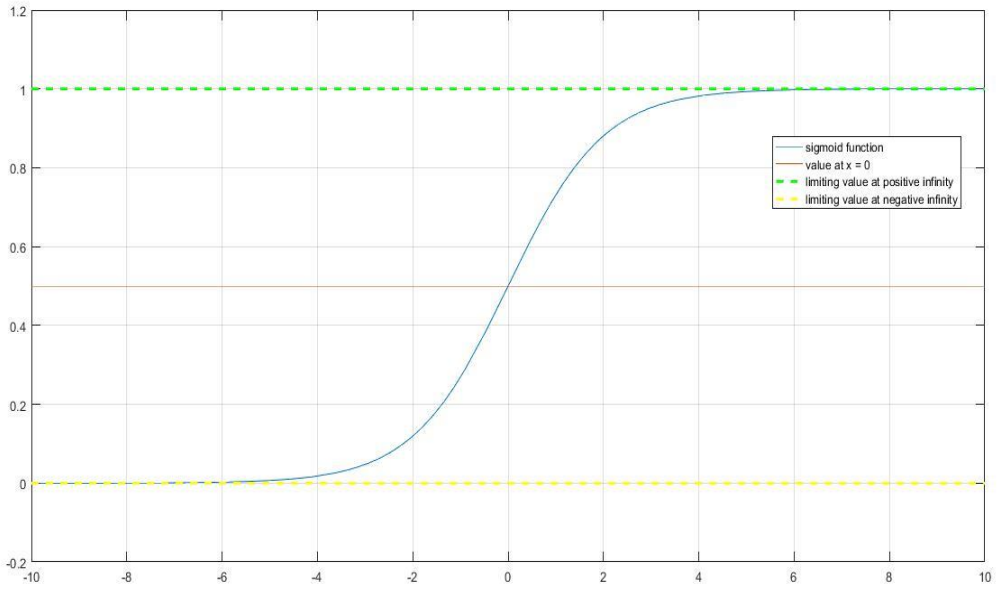

Fig. 2.1: Sigmoid Function

This method can also be used in classification tasks. It produces probability to make a binary choice. The target variable is either 0 or 1.

The idea behind this model is based on linear regression. 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 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 like predicting the win percentage of a team in a football match, the likelihood of rain on a specific date, diagnosis of diseases, etc.

2.2.3 Support Vector Machine

Support Vector Machine is a supervised learning algorithm which is effective for solving binary classification problems due to its strength in high-dimensional spaces. It works by finding the optimal hyperplane that best separates the data points of different classes.

A hyperplane is a decision boundary that separates different classes in the feature space. SVM aims to maximize the margin, which is the distance between the hyperplane and the closest data points from 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 in two regions for binary problems. Some of the common kernel functions are:

- **Linear Kernel:** Used when data is linearly separable.
- **Polynomial Kernel:** Data is mapped into polynomial space.
- **Radial Basis Function (RBF):** Works well for complex, non-linear data.

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

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

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

The hinge loss ensures 0 loss when SVM correctly classifies the data points and for each misclassified points, a penalty is added.

2.2.4 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 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 each tree. Each tree learns from a different subset of data in order to reduce variance and overfitting. These decision trees are built using a subset of features and they split data based on the most informative feature at each step. Their predictions are combined for final output.

Random Forest does not use a single loss function like SVM. Instead, it minimizes errors through averaging predictions from multiple trees.

2.2.5 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, ANN is widely used in fields like image recognition, Natural Language Processing (NLP), medical diagnosis, financial forecasting, etc.

ANN consists of three main types of layers:

- **Input Layer:** Receives raw data and assigns a weight to each input before passing to the next layer.
- **Output Layer:** Produces the final prediction and uses an activation function. ReLU (Rectified Linear Unit), Sigmoid, Tanh, etc. are some of the popular activation functions.
- **Hidden Layers:** Perform computations by applying weights, biases, and activation functions. Deep Neural Network is formed by multiple hidden layers. Each neuron computes:

$$\boldsymbol{z = W * X + B}$$

Where,

W = weight matrix

X = input

B = bias

The following table provides the computational efficiency (time complexity) of different machine learning methods (Gupta, 2023, Reddy, 2023):

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	$O(i * n * p)$	$O(p)$

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

3. ATTRIBUTES

The database we used contains information of 3276 water bodies including lakes, rivers and oceans (Dataset of water potability 2020). It contains 9 features about 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: Measure of water clarity

Potability: 1 indicates safe water, 0 indicates unsafe

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

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

4. RESULTS AND ANALYSIS

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

In case of KNN model, $k = 1$ and $k = 3$ led to the accuracies of 67.14% and 58.57% respectively. Then the accuracy increased gradually and reached a peak accuracy of 71.43% for k = 5. Then the accuracies decreased and stayed around 62% in the long run. Primarily, the high accuracy for smaller k-values was due to overfitting.

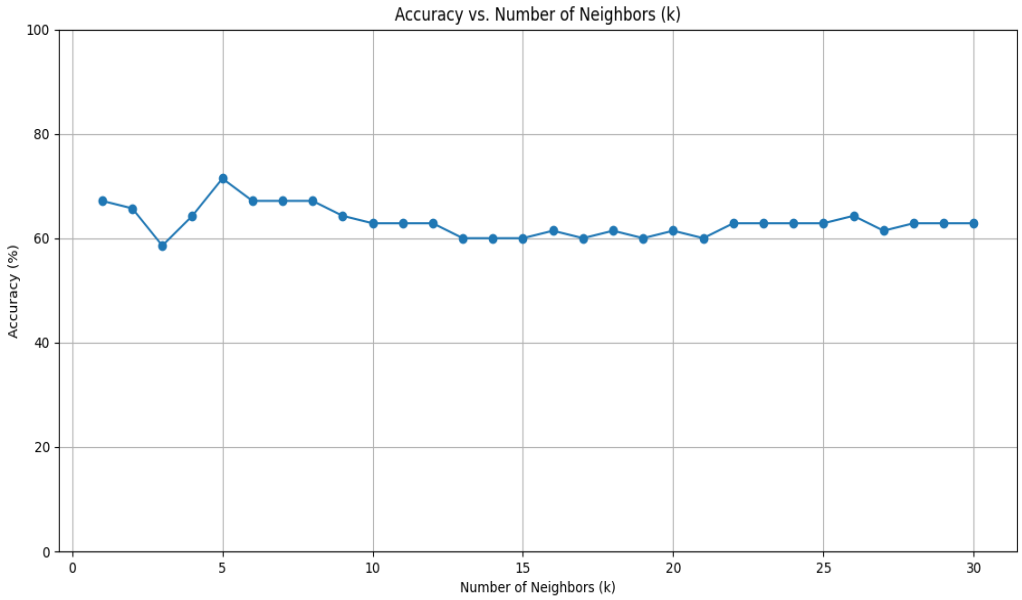

Fig 4.1: Graph of accuracy in KNN method for k-value up to 30. The accuracies slightly increased for k = 4 and k = 5, then began to decrease while being more than 60% always. (95% Confidence Interval for Accuracy: (55.5%, 62.8%))

The modified KNN model gave us an accuracy of 60% for $k = 1$ and 64.29% for $k = 3$. At first, it looks like this modified algorithm did not improve the accuracy at all. However, if we check the accuracy for the value of k up to 30, we see the performance being improved greatly.

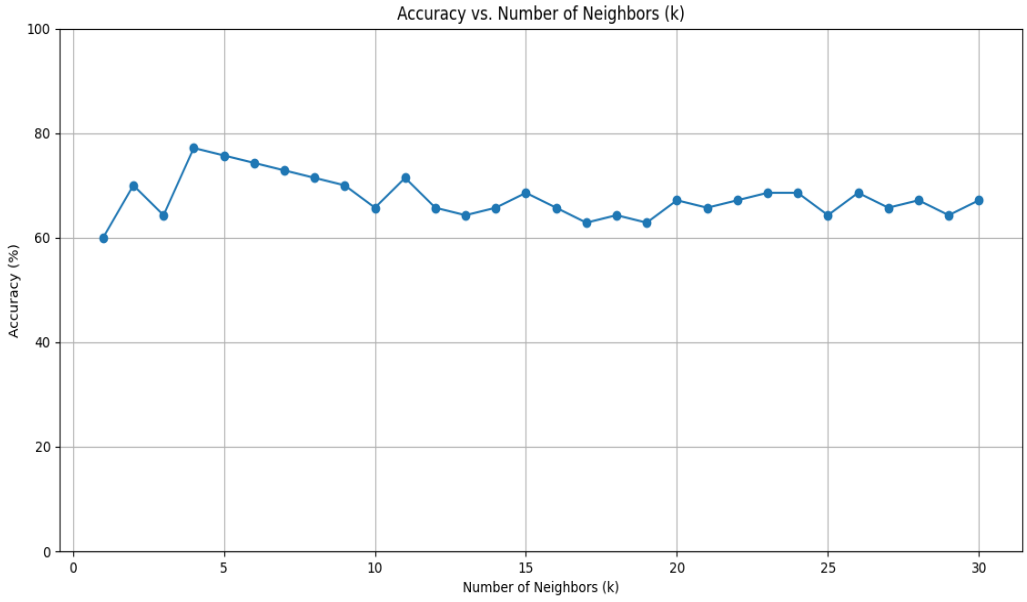

Fig 4.2: Graph of accuracy in the Modified KNN method for k-value up to 30. For most k-values, the accuracy was above 67% which makes it better than the other models. (95% Confidence Interval for Accuracy: (62.8%, 70.3%))

This new model achieved 70% accurate result at $k = 9$. For majority of the k-values, the accuracy remains above 67% and the peak accuracy is gained at $k = 4$ which is 77.14%. After that, the accuracy slightly fell off but as the value of k keeps getting bigger, the accuracy reaches a steady level around 67%.

The following table shows the accuracies of the methods we have discussed throughout our paper. We can clearly observe that our modification of KNN model is better at predicting water potability than all the other models.

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)

We also used 10-fold cross validation technique to check whether the improved accuracy of the modified model is significantly better than rest of the models. Also, there is a slight chance of biasness 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 1 of the folds as test data and the remaining 9 as training data. This reduces the chance of overfitting and gives us a much stable accuracy result. To perform this, we compared the 10 accuracies of the modified model with 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 was negative, our modified model is worse. If the p-value is less than 0.05, the difference in the accuracies is significant. The mean accuracy of the modified model is 66.57% and the confidence interval is (62.8%, 70.3%). The following table shows the mean accuracy of the 10 folds of each model (except the modified model) along with t-statistic and p-value, the confidence interval (95%) and a comment on whether the modified model is significantly better or not:

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%)

We attempted to remove all possible randomness associated with each model and train the model on the entire training dataset. However, there is a slight inherent randomness with the ANN model, so we checked 8 different random states. Our modified model was better than all of those models and in six of the cases, the accuracy was significantly better. So, we can confirm that the modification of the model is successful as the model is significantly better than rest of the models. Further change in this model can be made to improve accuracy by using different distance metrics.

5. CONCLUSIONS

We attempted to introduce a modification to the already existing KNN method by changing the distance metric and applying the idea of weighted distance to the model. Our purpose was building a model which can yield much better accuracy at predicting the purity of water than the methods mentioned before. The modified KNN model was significantly better than the KNN model in predicting water potability as there was a monumental increase in the accuracy. The modified KNN model also offered better accuracy than Logistic Regression, SVM, Random Forest and ANN model. Considering both accuracy and the fact that 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. As the future of machine learning looks promising, more research in this field will help us to build a model which has higher accuracy and precision. This study will aid in that journey by providing some insights about the algorithms in detecting water potability.

Author Contributions: Conceptualization, Zaman A.S.; methodology, Zaman A.S.; software, Fahim T.A.K. and Mahi H.M.; validation, Zaman A.S., Fahim T.A.K. and Mahi H.M.; formal analysis, Zaman A.S.; resources, Fahim T.A.K.; data curation, Fahim T.A.K.; writing—original draft preparation, Fahim T.A.K.; writing—review and editing, Zaman A.S. and

Mahi H.M.; supervision, Zaman A.S.; project administration, Zaman A.S.. 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.

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

REFERENCES

1. 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.
2. World Health Organization (WHO). (2022) *Drinking water.* Available at: https://www.who.int/news-room/fact-sheets/detail/drinking-water (Accessed: 11 November 2024).
3. Clark, R.M. and Hakim, S., 2014. Securing water and wastewater systems. *Cham: Springer International, 10,* pp.978-3.
4. 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.
5. 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.
6. UN Water (2024) *Progress on ambient water quality – 2024.* Available at: https://www.unwater.org/publications/progress-ambient-water-quality-2024-update (Accessed: 14 November 2024).
7. 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.
8. 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.
9. 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.
10. 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.
11. Kaddoura, S., 2022. Evaluation of machine learning algorithm on drinking water quality for better sustainability. *Sustainability, 14(18),* p.11478.
12. 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 Synthetic Minority Oversampling Technique and Explainable AI. *Computational Intelligence and Neuroscience, 2022(1),* p.9283293.
13. 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.

14. 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.
15. 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.
16. GeeksforGeeks (2024) *Supervised Learning*. Available at: https://www.geeksforgeeks.org/supervised-unsupervised-learning/ (Accessed: 14 November 2024).
17. Shalev-Shwartz, S. and Ben-David, S., 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.
18. GeeksforGeeks (2024) *K Nearest Neighbor*. Available at: https://www.geeksforgeeks.org/k-nearest-neighbours/ (Accessed: 14 November 2024).
19. Gupta, R. (2023) *Time complexity in machine learning.* Available at: https://medium.com/@riteshgupta.ai/time-complexity-in-machine-learning-4c253919e871 (Accessed: 3 March 2025).
20. Reddy, A. (2023) *Machine learning*. Available at: https://anudeepareddy-s.medium.com/machine-learning-c85118ec9a1b (Accessed: 3 March 2025).
21. Kaggle (2024) *Dataset of Water Potability*. Available at: https://www.kaggle.com/datasets/adityakadiwal/water-potability (Accessed: 14 November 2024).