2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC)

Water Quality Drinking Classification Using Machine Learning

Gasbaoui Mohammed el Amin, Benkrama Soumia, Bendjima Mostefa

Laboratory of TIT
Mathematics and Computer Science Department
Faculty of Exact Sciences
Tahri Mohammed University, Bechar, Algeria

gasbaoui.mohammedelamin@univ-bechar.dz
benkrama.soumia@univ-bechar.dz
bendjima.mostefa@univ-bechar.dz

2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC) | 979-8-3503-4974-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICEEAC61226.2024.10576539

***Abstract*—Water is crucial for all known forms of life. It is required for the biochemical processes that occur within living organisms. Human beings need water to survive. It helps maintain bodily functions, regulate body temperature, and transport nutrients and oxygen to cells. Drinking an adequate amount of water is essential for good health. It helps in digestion, circulation, absorption of nutrients, and the removal of waste products from the body. In this paper, we conduct an analysis and exploration of the nine parameters that are essential for assessing the quality of drinking water. Additionally, we compare the performance accuracy of five different machine learning algorithms including (Random Forest RF, k-Nearest Neighbors KNN, Support Vector Machine SVM, XGBoost, and Decision Tree DT). Our results indicate the Support Vector Machine (SVM) achieved the highest accuracy with a score of 70%, outperforming the other algorithms. Additionally, we discuss and analyze the key factors influencing model accuracy. To engage more with the deployed model, we provide a front-end dashboard for collecting user data input, which is connected to a pre-trained model through a RESTful API.**

***Keywords*— water features, machine learning, drinking water quality, data analysis, model accuracy**

I. INTRODUCTION

Water covers over two-thirds of the Earth's surface, making it a vital resource for sustaining life. Despite its abundance, the available supply of usable water is finite. Water holds a pivotal role on our planet, yet its ubiquitous presence often leads us to overlook its fundamental significance in our daily existence [1]. Safe and easily accessible water is essential for public health, serving various purposes such as drinking, household needs, food production, and recreation. Enhancing water supplies and implementing better water resource management can contribute to the well-being of nations, fostering prosperity and alleviating poverty [2]. The quality of water significantly impacts our daily lives and is a paramount concern in urban management. Conventional approaches to urban water quality management predominantly revolved around regular quality indicator assessments, encompassing physical, chemical, and biological parameters. Yet, the delayed detection of biological indicators has escalated health risks, resulting in incidents like widespread infections in numerous large cities [3]. Water drinking quality refers to the characteristics and conditions of water that make it safe and suitable for human consumption. It encompasses various factors and standards that ensure the water is free from contaminants, pathogens, and harmful substances, making it safe for people to drink without posing health risks. Key aspects of water drinking quality include the absence of harmful bacteria, viruses, and parasites, compliance with chemical and physical parameters within established guidelines and the absence of any unpleasant tastes, odors, or discoloration that could make the water unpalatable or unsafe to drink.

As the volume of data related to the aquatic environment continues to grow rapidly, machine learning has emerged as a crucial tool for tasks such as data analysis, classification, and prediction. Unlike conventional models commonly employed in water-related studies, machine learning-driven models excel at tackling intricate nonlinear challenges efficiently. Within the realm of water environment research, machine learning has found applications in areas such as designing, monitoring, simulating, assessing, and enhancing diverse water treatment and management systems [4]. The remainder of this paper is organized as follows: In the second section, we review related work in this research area. The third section focuses on the analysis and exploration of the water quality dataset, where we apply five different machine learning algorithms: Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and XGBoost. Additionally, we create a userfriendly dashboard for data input. The fourth section is dedicated to discussing our findings and presenting the results. Finally, in the last section, we offer conclusions drawn from our work and

979-8-3503-4974-0/24/$31.00 ©2024 IEEE

discuss potential future directions and aspirations for further research.

II. Related Works

Machine learning has found extensive use within the domain of drinking water treatment and management systems, spanning various aspects such as the management of drinking water sources, the optimization of treatment processes, the enhancement of water distribution systems, and the facilitation of decision-making. [4] The author created and implemented an integrated intelligent water quality monitoring system for aquaculture. This system employed sensors to continuously gather real-time data on parameters such as temperature, pH, dissolved oxygen, and turbidity in aquaculture water. The assessment of water quality was conducted using the water quality index method.

In this study [5] The authors introduced an innovative system designed to enhance the efficiency of water quality monitoring, aligning with the ongoing pollution control initiatives. In this context, a well-organized methodology was employed to gather quality parameters, with a particular focus on chemical indicators. The collected data was seamlessly transmitted to the cloud in real-time, enabling continuous monitoring of water quality and providing immediate access to real-time data for various chemical and biological indicators, including pH, dissolved oxygen, total dissolved solids, turbidity, and more.

In [6] this research delves into the assessment of artificial intelligence methodologies, specifically artificial neural networks (ANN), Group Method of Data Handling (GMDH), and Support Vector Machines (SVM), in their capability to forecast water quality parameters within the context of Tireh River, situated in southwestern Iran. Various transfer and kernel functions were examined in the process. The findings from the analysis of ANN and SVM models collectively suggest their effectiveness in predicting water quality components.

The authors in [7] introduced two innovative hybrid machine learning models based on decision trees, aiming to enhance the precision of short-term water quality predictions. These hybrid models incorporate fundamental components, namely, extreme gradient boosting (XGBoost) and random forest (RF), each equipped with advanced data denoising techniques. The dataset used for this study comprises water resource information collected from the Gales Creek site in the Tualatin River, which is recognized as one of the most polluted rivers worldwide. The primary objective of these two hybrid models is to forecast six key water quality indicators, encompassing water temperature, dissolved oxygen, pH value, specific conductance, turbidity, and fluorescent dissolved organic matter. Notably, these hybrid models exhibit improved accuracy in short-term water quality prediction.

Wu et al.[3] presented a potential remedy in the form of a risk analysis framework designed for urban water supply systems. This framework leverages indicator data gathered from industrial processes to detect and respond to alterations in water quality, thereby identifying potential risks. They have introduced the Adaptive Frequency Analysis (Adp-FA) method to analyze this data, harnessing frequency domain information from indicators to uncover their interdependencies and make individual predictions. The outcomes of their research indicate that their approach outperforms existing methods in various respects. This suggests that it has the potential to effectively facilitate early warnings for industrial water quality risks and provide valuable decision support.

In our study, we analyze and explore water quality data features, comparing the accuracy of five different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost, and K-Nearest Neighbors (K-NN) for water quality drinking classification.

III. Experiments Part

We used a Kaggle dataset of water quality with 3276 different water bodies in rows and nine features (parameters) in columns. first, let's try to explore the different features.

PH value : PH serves as a crucial parameter for assessing the acid-base equilibrium of water, acting as an indicator of its acidic or alkaline properties. The World Health Organization (WHO) has established a recommended pH range, with a maximum permissible limit falling between 6.5 to 8.5 [8].

Hardness : Hardness in water primarily arises due to the presence of calcium and magnesium salts, which are leached from geological formations during the water's passage. The duration of contact between water and these mineral-rich materials plays a pivotal role in influencing the overall hardness level found in untreated water [8].

Solids (Total dissolved solids - TDS): Water exhibits the capacity to dissolve various inorganic and certain organic minerals and salts, encompassing potassium, calcium, sodium, bicarbonates, sulfates, and more. These minerals can impart undesirable taste and a diluted appearance to the water. Consequently, this parameter holds significant importance in determining water usability. A high TDS value indicates a heightened mineral content in the water. The recommended threshold for TDS in drinking water is 500 mg/l, with a maximum allowable limit of 1000 mg/l [8].

Chloramines: Chlorine and chloramine stand as the primary disinfection agents employed within public water supply systems. Chloramines typically originate from the introduction of ammonia into the water treatment process alongside chlorine for the purification of drinking water. In the context of drinking water safety, chlorine concentrations of up to 4 mg/l are generally regarded as acceptable [8].

Sulfate: Sulfates are naturally occurring compounds abundant in minerals, soil, and geological formations. They are pervasive in the environment, being found in the

979-8-3503-4974-0/24/$31.00 ©2024 IEEE

atmosphere, groundwater, plants, and food sources. The predominant industrial application of sulfates lies within the chemical sector. In the majority of freshwater sources, sulfate concentrations typically span the range of 3 to 30 mg/l [8].

Conductivity: In its pure form, water exhibits poor electrical conductivity, acting as an effective insulator. However, the presence of ions in water can significantly enhance its ability to conduct electric current. According to the standards set by the World Health Organization (WHO), the Electrical Conductivity (EC) value of water should not exceed 400 µS/cm [8].

Organic Carbon : Total Organic Carbon (TOC) in source waters originates from the decomposition of natural organic matter (NOM) and can also have synthetic sources. TOC serves as a metric for quantifying the collective carbon content within organic compounds present in pure water. In accordance with the guidelines established by the United States Environmental Protection Agency (US EPA), the acceptable limit for TOC in treated or drinking water is set at less than 2 mg/l [8].

Trihalomethanes : often abbreviated as THMs, are chemical compounds that can be detected in water subjected to chlorine treatment. The presence of THMs in drinking water is subject to variation, dependent on factors such as the organic content within the water, the quantity of chlorine used for treatment, and the water's temperature during the treatment process. Generally, THM concentrations of up to 80 ppm are deemed safe for consumption within drinking water [8].

Turbidity : is contingent upon the volume of solid particles suspended within water, the World Health Organization (WHO) has put a recommended value for turbidity, which stands at approximately 5.00 Nephelometric Turbidity Units (NTU) [8].

The table 1 summarizes the desirable range of conditions for certain features in drinking water.

TABLE 1
DESIRABLE CONDITIONS RANGE OF CERTAIN DRINKING WATER FEATURES

Features	Desirable Conditions
PH value	between 6.5 to 8.5
Total dissolved solids - TDS	between 500 to 1000 mg/L
Chloramines	up to 4 mg/L
Sulfate	Between 3 to 30 mg/L
Conductivity	Less than 400 µS/cm (siemens per cm)
Organic Carbon	less than 2 mg/L
Trihalomethanes	up to 80 ppm (parts per million)
Turbidity	Close to 5.00 NTU (Nephelometric Turbidity Units)

Table 2 shows some samples of the dataset drinking water quality, Our primary goal is to leverage machine learning algorithms to accurately predict the potability status of water based on a set of available features. Within our dataset, the 'potable' column serves as a crucial indicator: a value of 0 signifies non-potable water (not suitable for human drinking), whereas a value of 1 denotes potable water (safe for human consumption). By harnessing various machine learning techniques, we aim to develop robust models capable of reliably discerning the potability of water from its associated features.

TABLE 2
SOME SAMPLES DATASET

Features	Entry 1	Entry 2
Ph	NaN	3.716080
Hardness	204.890455	129.422921
Solids	20791.318981	18630.057858
Chloramines	7.300212	6.635246
Sulfate	368.516441	NaN
Conductivity	564.308654	592.885359
Organic carbon	10.379783	15.180013
Trihalomethanes	86.990970	56.329076
Turbidity	2.963135	4.500656
Potability	0	0

We used five different machine learning algorithms :

- Decision Tree
- XGBoost
- K-Neighbours
- SVM
- Random Forest

First, let's tackle the issue of missing values in datasets. Missing data can pose challenges for many machine learning algorithms. Therefore, it's essential to identify and replace missing values in each column of your input data before proceeding with your prediction modeling. Table 3 shows the missing value of some features.

TABLE 3
SOME FEATURES THAT HOLD MISSING VALUE

Features	Missing value
ph	491
Sulfate	781
Trihalomethanes	162
Other features	0

Replacing missing values with the mean is a widely employed data preprocessing technique, which entails substituting any absent values with the average value of the respective feature or column. The following snapshot shows how to replace the missing value with the mean of the feature (mean column)

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
```

979-8-3503-4974-0/24/$31.00 ©2024 IEEE

```
df['ph'] = df['ph'].fillna(df['ph'].mean())
df['Sulfate'] = df['Sulfate'].fillna(df['Sulfate'].mean())
df['Trihalomethanes'] = df['Trihalomethanes'].fillna(df['Trihalomethanes'].mean())
```

We use 80 % of data for training and 20 % for test with a constant random state . see the following snapshot.

```
X = df.drop('Potability', axis=1)
y = df['Potability']
# import StandardScaler to perform scaling
 from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X = scaler.fit_transform(X)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

A. Decision Tree Machine Learning

A decision tree in machine learning is a predictive model that uses a tree-like structure to make decisions or predictions based on input data. It is a supervised learning algorithm that is commonly used for both classification and regression tasks. In a decision tree, the dataset is split into subsets based on various attributes or features, and at each node of the tree, a decision is made about which subset to explore next. This process continues until a prediction or classification is made at the tree's terminal nodes, known as leaf nodes.

We use a decision tree with 8 *max depth* and a *stable random state*

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
model_dt=DecisionTreeClassifier( max_depth=8, random_state=42)
model_dt.fit(X_train,y_train)
pred_dt = model_dt.predict(X_test)
dt = accuracy_score(y_test, pred_dt)
print(dt)
print(classification_report(y_test,pred_dt))
```

B. K-Neighbours Machine Learning

K-Nearest Neighbors (K-NN): A machine learning algorithm that makes predictions by finding the *K* training examples in the dataset that are closest (most similar) to a new, unseen data point. We use a KNeighborsClassifier with 9 of *neighbors*.

```
from sklearn.neighbors import KNeighborsClassifier
model_kn = KNeighborsClassifier(n_neighbors=9)
model_kn.fit(X_train, y_train)
pred_kn = model_kn.predict(X_test)
kn = accuracy_score(y_test, pred_kn)
print(kn)
```

C. SVM Machine Learning

Support Vector Machine (SVM) is a powerful machine learning algorithm used for both classification and regression tasks. It's particularly effective in scenarios where the data points are not linearly separable. SVM is known for its ability to handle high-dimensional data and achieve good generalization performance.

```
from sklearn.svm import SVC, LinearSVC
model_svm = SVC(random_state = 42)
model_svm.fit(X_train, y_train)
pred_svm = model_svm.predict(X_test)
svm = accuracy_score(y_test, pred_svm)
print(svm)
```

D. Random Forest Machine Learning

Random Forest is a machine learning ensemble algorithm that builds multiple decision trees during training and combines their predictions to improve accuracy and reduce overfitting. It works by randomly selecting subsets of the training data and features, creating a collection of decision trees. Random Forest is robust, handles high-dimensional data well, and is less prone to overfitting compared to individual decision trees, making it a popular choice in various machine learning application.

```
from sklearn.ensemble import RandomForestClassifier
model_rf = RandomForestClassifier( random_state=42)
model_rf.fit(X_train, y_train)
pred_rf = model_rf.predict(X_test)
rf = accuracy_score(y_test, pred_rf)
print(rf)
```

E. XGBoost Machine Learning

Extreme Gradient Boosting is a machine learning algorithm that belongs to the gradient boosting family, specifically designed to optimize predictive model performance. XGBoost builds an ensemble of decision trees sequentially, where each tree corrects the errors of the previous ones. It employs a gradient descent optimization technique to find the optimal tree structure and weights for combining predictions. XGBoost is known for its speed, accuracy, and ability to handle complex relationships in data. We set the *max_depth*= 8 and *learning_rate*= 0.03.

```
from xgboost import XGBClassifier
model_xgb = XGBClassifier(max_depth= 8,random_state= 42,learning_rate= 0.03)
model_xgb.fit(X_train, y_train)
pred_xgb = model_xgb.predict(X_test)
xgb = accuracy_score(y_test, pred_xgb)
print(xgb)
```

F. Front-end dashboard

We deployed the SVM model after we saved it using the Flask framework for our web application because it yields

979-8-3503-4974-0/24/$31.00 ©2024 IEEE

better results. We exposed the classification service as a RESTful API that accepts GET requests. A RESTful API is a set of rules and conventions for facilitating communication between different software applications over the internet or a network. In Figure 1, you can see a simple web page that collects user input for nine parameters. These parameters are transmitted to the model via an HTTP GET request, and the model responds with a string indicating whether the water is potable or non-potable. Below is a snapshot illustrating the format of the HTTP GET request, which includes the nine parameters in the URL.

```
http://localhost:5000/predict?ph=3&hardness=204
&solids=20791&chloramines=7&sulfate=310
&conductivity=418&oc=10&trihalomethanes=66
&turbidity=2
```

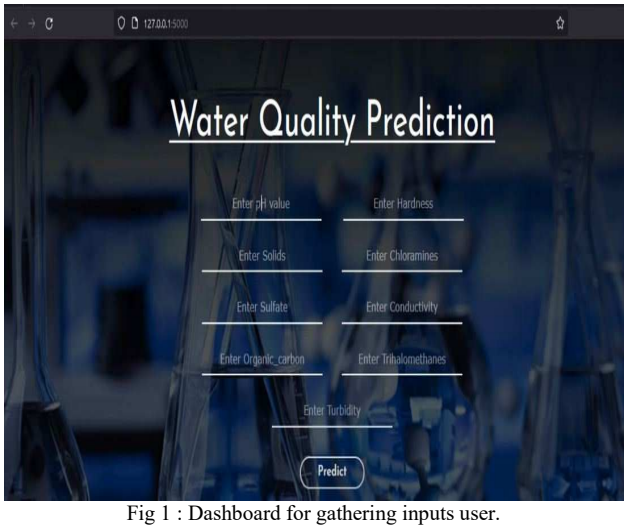

Fig 1 : Dashboard for gathering inputs user.

IV. RESULTS AND DISCUSSION

Table 4 presents a comparison of various algorithms, indicating that SVM outperformed others in terms of precision, recall, and F1-score for both classes (potable and not potable).

TABLE 4 : RESULTS OF PRECISION, RECALL, AND F1-SCORE OF FIVE MACHINE LEARNING ALGORITHMS

Algorithms	Status	Precision	Recall	F1-Score
DT	0 (Not potable)	0.68	0.83	0.74
	1 (Potable)	0.53	0.32	0.41
KNN	0 (Not potable)	0.68	0.82	0.74
	1 (Potable)	0.53	0.32	0.42
SVM	0 (Not potable)	**0.70**	**0.92**	**0.79**
	1 (Potable)	**0.69**	**0.32**	**0.44**
RF	0 (Not potable)	0.70	0.86	0.77
	1 (Potable)	0.61	0.32	0.44
XGBoost	0 (Not potable)	0.69	0.91	0.79
	1 (Potable)	0.67	0.32	0.44

Figure 2 illustrates that SVM achieved the highest accuracy rate of 70%, surpassing other methods. The achieved accuracy is influenced by the presence of imbalanced data. This situation is common in machine learning, where the distribution of classes in the dataset is uneven or imbalanced. In such cases, one class, often referred to as the minority class, may have significantly fewer examples than another class, known as the majority class. This class imbalance can introduce challenges during model training and evaluation. In the dataset we used, there were more instances of the 'not potable' class, accounting for 61% of label 0, while the 'potable' class represented 35% of label 1. Figure 3 illustrates the comparison between these two classes.

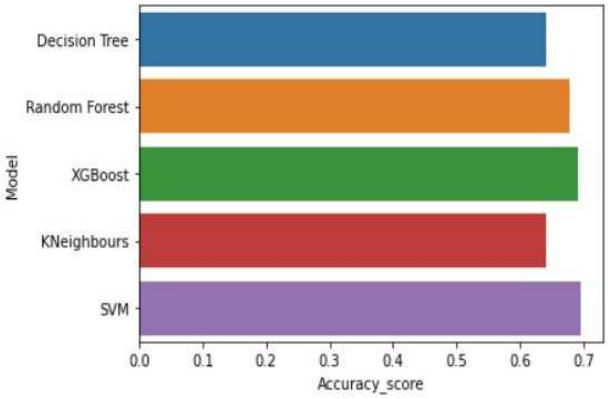

Fig 2 : Accuracy comparison of five machine learning algorithms.

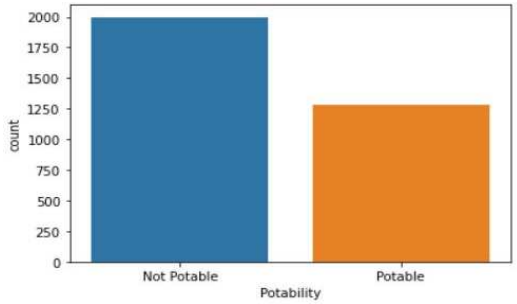

Fig 3 : Comparison instances between 'potable' and 'not potable' classes.

Dimensionality reduction is a technique employed to trim down the number of features in a dataset, all while retaining the most crucial information. Essentially, it involves transforming high-dimensional data into a lower-dimensional space that retains the core aspects of the original data. In the field of machine learning, high-dimensional data pertains to datasets characterized by a substantial abundance of features or variables. The presence of high-dimensional data can exacerbate the risk of overfitting, a situation in which the model becomes excessively tailored to the training data, hindering its ability to generalize effectively to new, unseen data.

In our case, we couldn't perform dimensionality reduction because there is no significant correlation between the

979-8-3503-4974-0/24/$31.00 ©2024 IEEE

features. Therefore, we utilized all nine features, Figure 4 shows the correlation between the features.

Outliers are data points that significantly differ from the majority of the data in a dataset. These data points are often unusual, rare, or distinct in some way, and they can have a substantial impact on the analysis and modeling of the data. Figure 5 illustrates that the distribution of the 'solids' feature significantly deviates from the other features. Removing outliers from this feature is not a viable option since it contains the majority of data points for the 'not potable' class. If we remove outliers from the 'solids' feature, it would result in most instances being categorized as 'potable' instead.

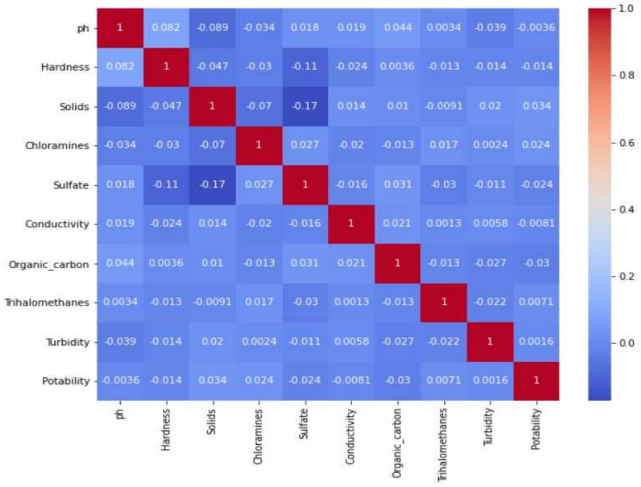

Fig 4 : Comparison the correlation between the features dataset.

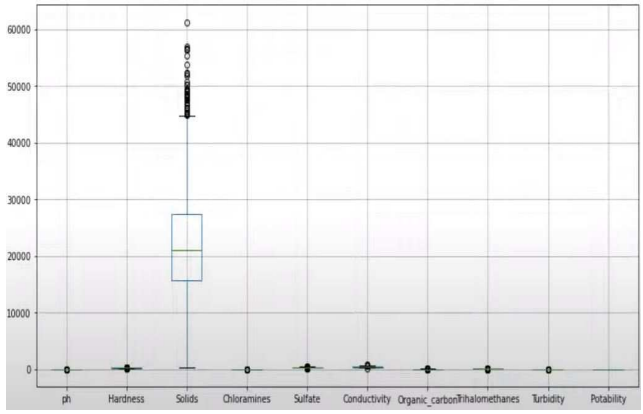

Fig 5: The distribution of the 'solids' feature compared to the others

V. Conclusion

This study examined the machine learning performance of various approaches, including XGBoost, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees (DT), in predicting components within a water quality dataset. To achieve this goal, the study gathered data from well-known datasets, encompassing variables such as pH, hardness, solids, sulfate, and conductivity . The findings indicated that Support Vector Machine (SVM) demonstrated the highest performance with an accuracy of 70%. Furthermore, the results underscored the significant influence of imbalanced data and the choice of imputation technique for addressing missing values and outliers on the ultimate accuracy of the model.

For the perspective and future work, we recommend an embedded system that measures the physical parameters like sulfate and trihalomethanes of the water using sensors and IoT context for rapid protection and early warning systems that can help authorities and water treatment plants take proactive measures to safeguard public health. exploring and analyzing advanced techniques that have an effect on the model precision and focus more on the data preparation and normalization like balancing the data, dealing with the missing values and outliers features. Considering multi-source data fusion techniques are an essential which is defined as an integration of data from various sources, such as sensor data, satellite imagery, weather data, and historical water quality data, to create comprehensive models for predicting water quality fluctuations.

References

[1] S. Kaddoura, 'Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability', *Sustainability*, vol. 14, no. 18, p. 11478, Sep. 2022, doi: 10.3390/su141811478.

[2] J. Patel *et al.*, 'A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI', *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–15, Sep. 2022, doi: 10.1155/2022/9283293.

[3] D. Wu, H. Wang, H. Mohammed, and R. Seidu, 'Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception', *IEEE Trans. Sustain. Comput.*, vol. 5, no. 3, pp. 377–388, Jul. 2020, doi: 10.1109/TSUSC.2019.2929953.

[4] Y. He, 'Design of water quality monitoring platform based on Embedded System', *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 631, no. 1, p. 012020, Jan. 2021, doi: 10.1088/1755-1315/631/1/012020.

[5] S. K. Vasudevan and B. Baskaran, 'An improved real-time water quality monitoring embedded system with IoT on unmanned surface vehicle', *Ecological Informatics*, vol. 65, p. 101421, Nov. 2021, doi: 10.1016/j.ecoinf.2021.101421.

[6] A. H. Haghiabi, A. H. Nasrolahi, and A. Parsaie, 'Water quality prediction using machine learning methods', *Water Quality Research Journal*, vol. 53, no. 1, pp. 3–13, Feb. 2018, doi: 10.2166/wqrj.2018.025.

[7] H. Lu and X. Ma, 'Hybrid decision tree-based machine learning models for short-term water quality prediction', *Chemosphere*, vol. 249, p. 126169, Jun. 2020, doi: 10.1016/j.chemosphere.2020.126169.

[8] 'Water Quality'. https://www.kaggle.com/datasets/adityakadiwal/water-potability (accessed Sep. 18, 2023).

979-8-3503-4974-0/24/$31.00 ©2024 IEEE