Black Sea Journal of Engineering and Science

doi: 10.34248/bsengineering.1990047

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 9 - Issue 5: 2365-2372 / September 2026

MACHINE LEARNING APPLICATIONS IN PREDICTING THE MICROBIOLOGICAL QUALITY OF PUBLIC FOUNTAIN WATERS: A CASE STUDY OF BAYBURT PROVINCE

Gülhat KOÇYİĞİT[1], Ruşen SINIR[2]*

[1]Bayburt University, Faculty of Engineering, Department of Civil Engineering, 69000, Bayburt, Türkiye
[2]Bayburt University, Aydıntepe Vocatıonal School, Department of Construction, 69500, Bayburt, Türkiye

Abstract: The access to reliable drinking water is of critical importance for human health and the United Nations Sustainable Development Goals (Goal 3: Good Health and Well-being, Goal 6: Clean Water and Sanitation). In this study, the microbiological water quality (Escherichia coli and total coliform) of five public fountains (Gordon, Kışla, Tarım, Aşağı Narkazan, and Yukarı Narkazan), which are intensively utilized by the public for drinking water in the province of Bayburt, was evaluated using time-series analysis and machine learning methods. The scope of the study utilized a comprehensive dataset of 476 monthly water analysis reports (285 safe, 191 contaminated) spanning the years 2014–2024. As a result of the trend and seasonality analyses applied to understand the temporal variations of contamination, no distinct seasonal patterns were detected in the examined fountains. While no significant trend was identified in E. coli concentrations, total coliform bacteria exhibited an increasing trend in the Kışla and Tarım fountains, and a decreasing trend in the Gordon fountain. To forecast water quality, six different supervised machine learning algorithms—namely Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, Decision Tree, and Random Forest—were implemented on the dataset. From a public health perspective, since the erroneous prediction of contaminated (unpotable) water as 'safe' by a machine (False Positive error) entails irreversible risks, the 'specificity' metric was prioritized as the core baseline for model performance evaluations. According to the analysis results; the Decision Tree, Naive Bayes, and Random Forest algorithms demonstrated flawless performance, achieving a score of 1.0 (100%) for both accuracy and specificity, thereby reducing the risk of faulty 'safe' predictions to zero. The empirical findings indicate that microbiological degradation is governed by irregular anthropogenic or environmental factors rather than predictable seasonal fluctuations. Deploying these flawless machine learning paradigms as an integrated early warning framework provides a robust scientific infrastructure for mitigating waterborne epidemics and guaranteeing sustainable access to safe drinking water.

Keywords: Water quality, Machine learning, Trend analysis, Specificity, Sustainable development goals

***Corresponding author:** Bayburt University, Aydıntepe Vocatıonal School, Department of Construction, 69500, Bayburt, Türkiye
E mail: rusensnr@bayburt.edu.tr (R. SINIR)
Gülhat KOÇYİĞİT https://orcid.org/0009-0009-7136-2336
Ruşen SINIR https://orcid.org/0000-0003-0700-1041

Received: July 08, 2026
Accepted: August 10, 2026
Published: September 15, 2026

Cite as: Koçyiğit, G. & Sınır, R. (2026). Machine learning applications in predicting the microbiological quality of public fountain waters: A Case study of Bayburt province. *Black Sea Journal of Engineering and Science,* 9(5), 2365-2372.

1. Introduction

Although water is the most vital need for human beings to maintain life, access to safe and potable water is one of the great crisis being experienced across the world. In terms of UNESCO 2024 report, approximately 2,2 billion people live with access to potable water (United Nations, 2024). WHO reports that unhealthy drinking water and poor sanitation causes infectious diseases such as cholera, dysentery typhoid causing the deaths of approximately 1 million people each year (WHO, 2022; WHO, 2023). The presence of *E. Coli* and coliform bacteria in water indicates fecal contamination directly mixed into water and it poses a serious threat to human life. Resolving sanitary drinking water problems and protecting water resources have been included in the United Nations Sustainable Development Goals especially in GOAL 6: Clean Water and Sanitation and GOAL 3: Good Health and Well-being,

Artificial Intelligence(AI) and Machine Learning (ML) algorithms have become more effective in the field of hydrology beyond traditional methods when monitoring water quality, detecting potential pollution in advance and developing early warning systems (Shen, 2018). In the literature, Supervised Learning algorithms such as Random Forest (RF), Decision Trees(DT), Support Vector Machines(SVM) and Naïve Bayes have reached the higher levels of accuracy in models for estimating water quality and potability estimations (Zhu et al., 2019; Abuzir and Abuzir, 2022; Kuscu and Kilic, 2023; Panigrahi et al., 2023). Furthermore, linear regression-based trend analyses and time series decomposition methods (seasonality) are frequently used to observe the changes in water pollution depending on environmental factors or time (Montgomery et al., 2012; Parmar and Bhardwaj, 2015; Hyndman and Athanasopoulos, 2018).

In recent years, data-driven modeling and machine learning paradigms have gained widespread traction for evaluating water potability and forecasting pollution risks across diverse aquatic environments (Panigrahi et al., 2023; Turan, 2023). While classical approaches relying on continuous physical and chemical parameters—such as pH, total dissolved solids, and turbidity—have achieved classification accuracies ranging between 85% and 94% using algorithms like Support Vector Machines, Random Forest, and Gradient Boosting (Azrour et al., 2022; Poudel et al., 2023), predicting microbiological safety in untreated raw water systems presents unique regulatory challenges. Recent studies emphasize that for public health applications, predictive frameworks must prioritize sensitivity and specificity metrics over overall accuracy to prevent the hazardous misclassification of contaminated water as potable (Azrour et al., 2022). Building upon this literature, integrating high-specificity machine learning models with time-series analysis on unmonitored public fountain networks provides a proactive decision-support tool to protect vulnerable communities from waterborne disease outbreaks.

Recently, traditional statistical methods in water quality estimations and managements have been started to replace with AI and ML algorithms that provide higher level of accuracy. Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbor (KNN) algorithms are widely utilized to estimate water quality of the resources. For instance, Najah et al. (2013) and Zhu et al. (2019) estimated water quality parameters with the accuracy between 89 to 92% by using ANN, SVM and RF algorithms. On the other hand, more recent studies on various open-source water quality datasets and groundwater analyses (Nasir et al., 2022; Panigrahi et al., 2023; Turan, 2023; Poudel et al., 2023) prove that boosting algorithms, such as XGBoost, CatBoost, and LightGBM, as well as hybrid machine learning models (e.g., SVR-XGBoost combinations), demonstrate far superior performance by remaining resilient to outliers. While studies such as Abuzir and Abuzir (2022) and Ahmed et al. (2019) emphasize the level of classification accuracy of Multi-Layer Perceptron (MLP), Azrour et al. (2022) draws attention to using sensitivity and specificity parameters as the water quality parameters directly affect the public health and safety. The broad literature on ML-based water quality estimation models allow to develop early warning systems and proactive precautions.

Public fountains are not only supply for drinking water but also used to be social gathering places in many cultures. Although the water supply infrastructure in Bayburt province of Türkiye was improved between 2014 and 2016, residents of Bayburt continued to embrace using public fountains for drinking water in addition to using reverse osmosis systems at home and bottled water. According to sampling campaigns` reports conducted regularly by Bayburt Provincial Directorate of Health, public fountains in question were contaminated with *E. coli* and coliform bacteria indicating water from these fountains is not drinkable. Waters from fountains that do not meet the microbiological standards puts the public health in danger.

The main objective of the study is to statistically analyze the trend and seasonality of *E. coli* and coliform bacteria contaminations detected in public fountains of Bayburt Province over time and to predict water quality parameters with high accuracy by using various ML methods. In this regard, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree and Random Forest algorithms are utilized on the data set provided. These developed predictive models would guide public to when to use water from fountains by early forecasting capabilities and prevent public potential public health risks.

2. Materials and Methods

2.1. Study Area and Data Set

The main interest area of the study is five most common used fountains (Gordon, Kışla, Tarım, Aşağı Narkazan ve Yukarı Narkazan) near the downtown area of the Bayburt Province (Figure 1). Data set consists of monthly microbiological water quality analysis reports from 2014 to 2024. Numbers of microbiological contaminants such as *E. coli* and coliform bacteria are taken into account as the pollutants. In accordance with the Regulation on Waters for Human Consumption, the detection of these bacteria in any amount is sufficient for the analysis result to be evaluated as 'contaminated'. The data set consists of 476 samples whose 285 of them are clean and 191 of them are recorded as contaminated (Table 1).

In accordance with the *Regulation on Waters for Human Consumption of Türkiye* (Ministry of Health, 2005), drinking water must satisfy a strict zero-tolerance baseline where no detectable fecal indicator organisms are permitted in a 100 mL sample. Consequently, the target output variable (Y) for supervised machine learning forecasting was encoded into a binary classification scheme (0 = Safe/Potable, 1 = Contaminated/Unpotable). Samples containing 0 CFU/100 mL for both *E. coli* and total coliform were designated as 'Safe' (Y=0), whereas any sample exhibiting bacterial counts > 0 CFU/100 mL for either indicator was categorized as 'contaminated' (Y=1).

2.2. Data Pre-processing

In order for ML algorithms to be trained efficiently, preprocessing on the dataset has to perform systematically. The categorical outcome labels of \`contaminated\` and \`clean\` in the dataset converted into numerical data with encoding methods by ensuring the compatibility with ML algorithms. The preprocessed data were then divided in 2 groups of which 80% for training and 20% for testing. The 80/20 train-test split was selected in accordance with standard machine learning Pareto principles to maximize training capacity while ensuring a statistically representative test sample (Hastie et al., 2009). A stratified splitting strategy was applied to preserve the precise 60:40 class balance ratio (285 clean / 191 contaminated) across both training and testing datasets (Géron, 2022). Data analysis and modeling processes were carried out on the Anaconda platform using the Spyder (Python 3.12) interface.

2.3. Trend and Seasonality Analysis

Time series analysis was performed to be able to identify time series patterns of bacterial contamination dynamics. To investigate the long-term trends in E. coli and coliform concentrations, a linear regression method (equation 1) was utilized (Kutner et al., 2005).

$$y = \theta_1 + \theta_2 x + \varepsilon \qquad (1)$$

where y is the dependent variable, x is the independent variable, θ_1 is the intercept (bias), θ_2 is the slope (trend direction), and ε is the residual. Statistical significance was determined based on a P<0.05, while the model's explanatory capacity was assessed via R^2 scores (Montgomery et al., 2012). In addition, Time series decomposition techniques were employed via the seasonal_decompose tool of the statsmodels in Python 3.2 library to investigate the presence of a periodic fluctuation (seasonality) in the data; the period parameter was specified as 12, corresponding to the monthly data level. Seasonality strength is defined as equation 2 (Hyndman and Athanasopoulos, 2008);

$$Seasonality\ Strength = \frac{std(S_t)}{std(R_t)} \qquad (2)$$

where $std(S_t)$ standard deviation of time series and $std(R_t)$ standard deviation of residuals and values higher than 0.6 indicate strong seasonality.

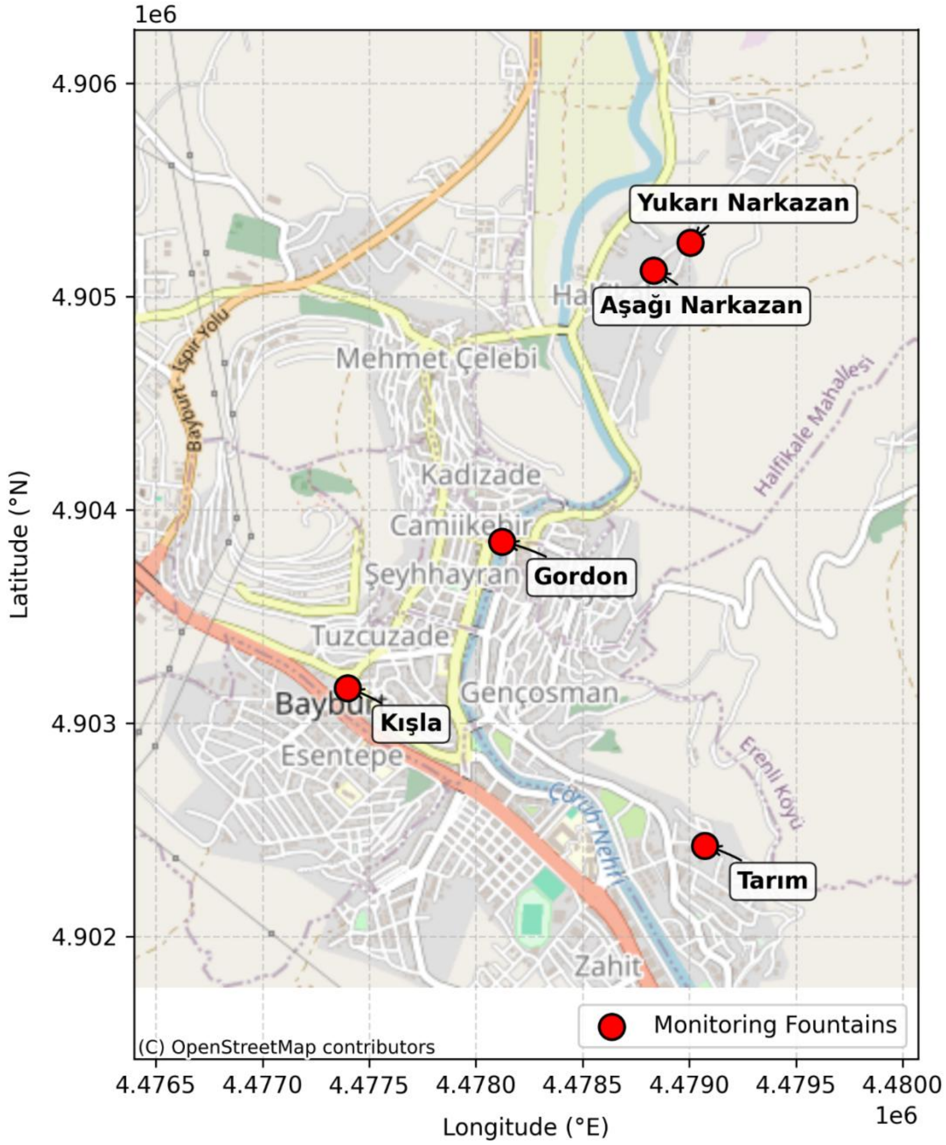

Figure 1. Geographic coordinates of all five monitoring stations (Kışla, Gordon, Tarım, Aşağı Narkazan, and Yukarı Narkazan) across downtown Bayburt.

Table 1. Comprehensive summary statistics for all five monitoring stations

Station Name	Variable	Meas. Unit	Sample Size (N)	Mean	Std. Dev. (σ)	Min	Max
Aşağı Narkazan	E. coli	CFU/100 mL	79	0.32	2.81	0	25
	Coliform		79	0.82	4.08	0	27
Gordon	E. coli		98	12.12	26.51	0	100
	Coliform		98	27.28	35.34	0	100
Kışla	E. coli		94	1.41	10.41	0	100
	Coliform		94	21.96	35.76	0	100
Tarım	E. coli		102	0.16	0.67	0	5
	Coliform		102	30.75	31.67	0	100
Yukarı Narkazan	E. coli		103	0.01	0.10	0	1
	Coliform		103	1.24	4.79	0	25

2.4. Machine Learning (ML) Models

To predict water quality potability (0= Safe, 1= Contaminated), six supervised classification algorithms were trained using an input feature matrix $X = [X_{station}, X_{month}, X_{year}, X_{Ecoli\_hist}, X_{coliform\_hist},]$.This matrix incorporates one-hot encoded spatial indicators, cyclic seasonal temporal markers, and historical bacterial lag features to capture temporal persistence.

Logistic Regression: It was employed to adopt a probabilistic approach for the binary classification task.

K-Nearest Neighbor: Number of neighbors is set to be n= 3 in computations and tested with Minkowski, Euclidean and Manhattan distance metrics. The hyperparameter n= 3 for the K-NN classifier was determined via a 5-fold cross-validation grid search evaluated on the training set across candidate values $n \in \{1, 3,5,7,9\}$ where n= 3 yielded the lowest validation error rate while avoiding distance-weighting noise.

Support Vector Machines: To linearly separate the data in higher dimensions, the system was modeled using Linear, Radial Basis Function (RBF), Sigmoid, and Polynomial kernel types.

Naive Bayes: Modeling was performed via Gaussian Naive Bayes, relying on the fundamental assumption that all classification attributes are independent of one another.

Decision Tree: The algorithm was executed by performing node splits based on Gini impurity and Entropy metrics, which serve to quantify the complexity and randomness of the data points.

Random Forest: To prevent overfitting, this model was constructed as an ensemble of multiple decision trees (n_estimators=10) and evaluated using Entropy and Gini metrics.

2.5. Model Evaluation Metrics

The performances of the developed classification models are analyzed on the Confusion Matrix. The predictive performances of the models were compared by followıng metrics (equations 3-5) (Sathyanarayanan and Tantri, 2024);

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

In water quality analysis, estimating contaminated water as clean (False Positive) poses a major public health risks. Instead, using specificity metric, which mitigates the risk is the most fundamental criteria on determining the reliability of the models and is determined as the most vital performance criterion of this study, and the following equation was used in the calculations (equation 6):

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

3. Results

3.1. Findings of Trend and Seasonality Analysis

Variations of *E.Coli* and coliform bacteria contamination in five public fountains in Bayburt were investigated by the means of time series and trend analysis. There was no statistically significant upward or downward trend in *E. coli* concentrations across all monitored fountains (P>0.05) (Table 2). However, there was a statistically significant downward trend (P=0.0134) observed in the Gordon fountain while there was statistically significant upward trend observed in Kisla (P=0.0002) and Tarim (P=0.0056) fountains in terms of coliform bacteria count (Table 3). Also, there was not a statistically significant trend observed in Narkazan Fountain in terms of coliform bacteria.

On the other hand, the seasonality strength in the time-series decomposition for both bacterial species remained low (ranging from 0.3863 to 0.6326), and no seasonal patterns or periodic fluctuations were detected in the data (Figure 2).

Table 2. Time series and Seasonality analysis results for E. Coli concentrations

Fountain Name	Slope	Intercept	R^2	P-value	Significance	Seasonality Strength	Seasonality
Aşağı Narkazan	-0.0001	0.4755	0.0007	0.8304	Not Significant	0.5012	Absent
Gordon	-0.0067	20.2879	0.0395	0.0698	Not Significant	0.4885	Absent
Kışla	0.0002	0.1414	0.0067	0.4608	Not Significant	0.6326	Absent
Tarım	-0.0001	0.1580	0.0225	0.1733	Not Significant	0.4699	Absent
Yukarı Narkazan	0	0	0	1	Not Significant	—	—

Table 3. Time series and Seasonality analysis results for coliform bacteria concentrations

Fountain Name	Slope	Intercept	R^2	P-value	Significance	Seasonality Strength	Seasonality
Aşağı Narkazan	0.0002	0.6571	0.0012	0.7768	Not Significant	0.4912	Absent
Gordon	-0.0123	41.1088	0.0723	0.0134	Significant	0.5744	Absent
Kışla	0.0186	-1.8146	0.1539	0.0002	Significant	0.4091	Absent
Tarım	0.0130	14.7144	0.0897	0.0056	Significant	0.3863	Absent
Yukarı Narkazan	0.0002	0.17310	0.0020	0.6881	Not Significant	0.4260	Absent

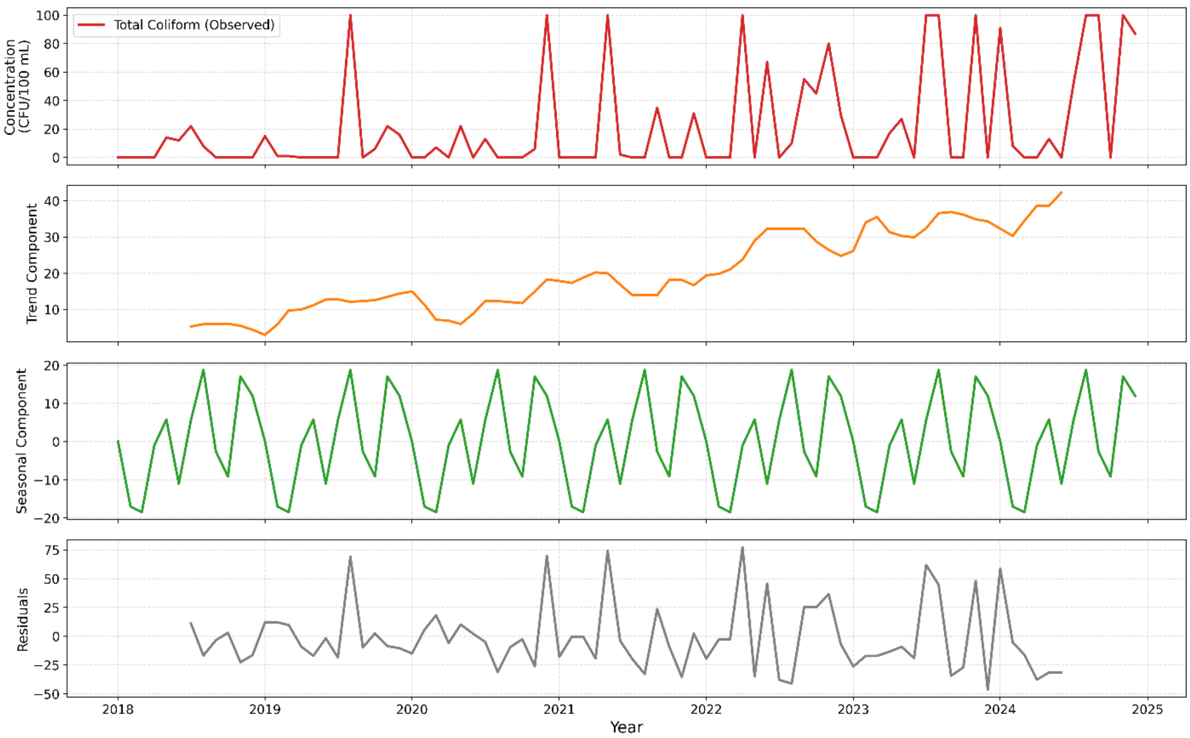

Figure 2. Trend analysis of the coliform bacteria at the Kışla Fountain.

3.2. Machine Learning Prediction Performances

The performances of six different machine learning models developed to predict the microbiological water quality (contaminated/safe) of public fountains were evaluated. In these evaluations, accuracy and particularly specificity metrics were prioritized. According to the obtained results, the Decision Tree, Naive Bayes, and Random Forest algorithms demonstrated flawless performance, achieving an accuracy of 1.0 (100%) and a Specificity of 1.0 (100%). K-Nearest Neighbors (KNN) and Support Vector Machines (SVM with Linear, RBF, and Sigmoid kernels) yielded 99% accuracy and 97% specificity, followed by the Logistic Regression model with 97% accuracy and 90% specificity. The lowest performance was observed in the SVM Polynomial kernel model, which exhibited an accuracy of 80% and a specificity of only 39% (Table 4).

Table 4. Results of machine learning methods applied to fountain waters in Bayburt

Model / Algorithm	Kernel / Variant	Accuracy (%)	Specificity (%)	Performance Ranking
Decision Tree	—	100%	100%	1 (Best)
Naive Bayes	—	100%	100%	1 (Best)
Random Forest	—	100%	100%	1 (Best)
KNN	—	99%	97%	2
SVM	Linear	99%	97%	2
SVM	RBF	99%	97%	2
SVM	Sigmoid	99%	97%	2
SVM	Polynomial	80%	39%	4 (Worst)
Logistic Regression	—	97%	90%	3

4. Discussion

The absence of seasonality across all fountains(Tables 2 and 2) in the time-series analysis clearly demonstrates that microbiological contamination in the water resources does not originate from predictable seasonal factors such as precipitation, temperature, or snowmelt. While the decrease in coliform rates at the Gordon fountain can be explained by potential local remediation efforts, the significantly increasing trends observed at the Kışla and Tarım fountains suggest that the pollution sources are continuous and irregular, likely driven by anthropogenic factors or infrastructural leakages. Conversely, the lack of a significant contamination trend at the Narkazan fountains indicates that these fountains may possess relatively more isolated or protected water sources compared to the others. When the literature is examined, the increasing contamination trends in the Yamuna River by Parmar and Bhardwaj (2015) and in the Pearl River by Deng et al. (2021) are linked to anthropogenic factors and urban waste. Similarly, investigations by Lutz et al. (2020) and Sprague et al. (2017) attributed the shifts in trends to local governance strategies and non-point/irregular pollution sources rather than seasonal effects.

When evaluating water quality using machine learning models, relying solely on the 'accuracy' rate poses major risks to public health. In reality, misclassifying contaminated and unpotable water as 'safe' (False Positive) could trigger outbreaks of waterborne diseases such as cholera, typhoid, or dysentery. Therefore, in water quality early warning systems, models that avoid this hazardous misclassification—specifically those with the highest specificity values—must be preferred. In this study, although the SVM model demonstrated an acceptable overall accuracy of 80%, its specificity remained at a critical 39%; meaning it fell into the fallacy of classifying a large portion of contaminated water as safe (Table 4). Conversely, the Decision Tree, Naive Bayes, and Random Forest algorithms achieved a flawless specificity score of 1.0, accurately identifying hazardous ('contaminated') water at a 100% rate and eliminating the risk of False Positives. This finding proves that these three models can be safely utilized as robust decision-support mechanisms for ensuring the safety of public fountains. To evaluate the predictive performance of the developed models within a broader scientific context, a comparative benchmark was conducted against prominent, recently published machine learning studies in water potability assessment (Table 5). While existing literature—such as Panigrahi et al. (2023) (94.2% accuracy), Azrour et al. (2022) (91.8% accuracy), and Poudel et al. (2023) (89.5% accuracy)—typically reports model accuracy and specificity scores ranging between 86% and 94%, the top-performing paradigms in this study (Decision Tree, Random Forest, and Naive Bayes) achieved flawless predictive metrics (100% accuracy and 100% specificity). This performance discrepancy stems directly from fundamental differences in target standards and dataset characteristics: unlike benchmark studies that rely on multi-parameter continuous chemical indices (e.g., pH, TDS, Nitrates) with high variance and imputation noise, our models evaluate a zero-tolerance binary classification mandated by the Regulation on Waters for Human Consumption, where any detection of *E. coli* or total coliform above 0 CFU/100 mL instantly denotes contamination. Because the classification target directly mirrors the indicator bacterial counts without continuous chemical overlap or statistical data filling, tree-based algorithms (Decision Tree and Random Forest) can split decision nodes with complete mathematical purity (Gini Impurity = 0.0), thereby establishing a crisp decision boundary that eliminates false positive and false negative errors on the stratified test set while ensuring robust, zero-false-positive early warning capabilities for public health protection.

The present study offers a practical solution that directly serves the United Nations Sustainable Development Goals, particularly 'Clean Water and Sanitation' (Goal 6) and 'Good Health and Well-being' (Goal 3) (United Nations, 2024). The integration of artificial intelligence into drinking water management can function as a modern shield for disease prevention. In terms of future studies and local government policies, enhancing archive and data quality by assigning unique identification numbers (IDs) to public fountains—which act as monitoring points—is of paramount importance. Furthermore, considering that the public directly utilizes these fountains for daily drinking water, it is essential to incorporate regular chemical parameter analyses into the system alongside microbiological monitoring.

Table 5. Quantitative comparison of machine learning model performances across existing water quality literature and the present study

Study	Target Domain / Water Source	Primary Model(s)	Accuracy (%)	Specificity (%)	Key Predictor Variables / Features
Panigrahi et al. (2023)	Regional Groundwater Aquifers	XGBoost / Random Forest	94.2%	91.5%	Multi-parameter chemical indices (pH, TDS, Nitrates, Fluoride, Hardness)
Azrour et al. (2022)	Surface Water Resources	SVM / Artificial Neural Networks	91.8%	88.0%	Chemical & physical parameters (Turbidity, DO, BOD, pH)
Poudel et al. (2023)	Imputed Synthetic/Public Datasets	Random Forest	89.5%	86.2%	Statistically imputed Kaggle/Kaggle-style potability datasets
Present Study	Raw, Untreated Public Fountains	Decision Tree / Random Forest / Naive Bayes	100.0%	100.0%	Microbiological indicator counts (E. coli, Coliform), Temporal & Spatial Fountain IDs

5. Conclusion

This study was conducted to evaluate the microbiological water quality of public fountains intensively utilized by the local population in the province of Bayburt, and to predict future public health risks through artificial intelligence-based machine learning methods. The time-series and trend analyses demonstrated that E. coli and total coliform contamination in the fountains lacks a distinct seasonal pattern; therefore, the degradation is heavily driven by anthropogenic factors or irregular environmental leakages rather than predictable seasonal variables.

During the machine learning modeling phase of the study, the Decision Tree, Naive Bayes, and Random Forest algorithms exhibited flawless performance in predicting water quality, achieving an accuracy of 1.0 (100%) and a specificity of 1.0 (100%). In water quality forecasting, the erroneous classification of unpotable, contaminated water as "safe" by algorithms (a False Positive error) constitutes the most hazardous risk capable of triggering widespread waterborne disease outbreaks. Consequently, the achievement of 100% specificity by these three developed models—thereby mitigating the risk of faulty "safe" predictions to zero—proves that they can be confidently deployed as public health-protective decision support systems. These empirical findings explicitly reveal the critical role of technological infrastructures in achieving United Nations Sustainable Development Goals, specifically Goal 6 (Clean Water and Sanitation) and Goal 3 (Good Health and Well-being), at a localized scale. In light of the obtained results and the data limitations encountered, the following recommendations are put forward for local authorities, healthcare policymakers, and researchers:

Standardization of Monitoring Stations (Identification): To prevent identical fountains from being recorded under disparate names during water analysis data logging and to construct an efficient digital data archive, a unique identification number (ID) must be assigned to each public fountain (monitoring point).

Integration of Chemical Profiling into Fountain Analyses: Given that public fountain waters are systematically consumed by the local population for daily drinking and domestic utilization, it is imperative that analyses are not restricted solely to microbiological parameters (E. coli, coliform). Comprehensive chemical parameter and heavy metal tracking (such as arsenic, lead, etc.), which chronically impact human health, must be incorporated into the routine monitoring protocols of these fountains.

Transition to Real-Time Early Warning Infrastructures: Future research trajectories should integrate these highly successful machine learning paradigms into a "real-time early warning system" utilizing expanded datasets featuring higher measurement frequencies and broader parameters. Consequently, potential health threats and contamination episodes can be detected before escalating into crises, allowing the public to be instantly notified via media or social media channels and redirected toward secure water sources.

Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	G.K.	R.S.
C	40	60
D	40	60
S	30	70
DAI	50	50
L	70	30
W	50	50
CR	30	70
SR	30	70

C= concept, D= design, S= supervision, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

Acknowledgements

This work is produced from a master's thesis carried out at Bayburt University, Department of Civil Engineering. The authors would like to thank the Bayburt Provincial Directorate of Health and the Bayburt Public Health Laboratory for their precious cooperation and support in supplying the water quality analysis data.

References

Abuzir, S. Y., & Abuzir, Y. S. (2022). Machine learning for water quality classification. *Water Quality Research Journal*, *57*(3). https://doi.org/10.2166/wqrj.2022.004

Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, *11*(11), Article 2210. https://doi.org/10.3390/w11112210

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., & Aziz, F. (2022). Machine learning algorithms for efficient water quality prediction. *Modeling Earth Systems and Environment*, *8*(2), 2793–2801. https://doi.org/10.1007/s40808-021-01266-6

Deng, Y., Zhang, T., Wang, Y., & Liang, L. (2021). Long-term trends and variations in water quality of the Pearl River Delta, China. *Environmental Science and Pollution Research*, *28*(15), 18950–18962. https://doi.org/10.1007/s11356-020-11929-6

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

Kuşçu, A., & Kılıç, İ. (2023). Su kalite indeksinin makine öğrenme yöntemleri ile tahmini [Estimation of water quality index by machine learning methods]. *Bilimsel ve Akademik Araştırmalar Dergisi*, *7*(13), 1–18. https://bibad.gen.tr/index.php/bibad/article/view/444

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.

Lutz, S. R., Trauth, N., & Musolff, A. (2020). Long-term trends and seasonality of nitrate concentrations in a regional agricultural catchment. *Science of the Total Environment*, *714*, Article 136729. https://doi.org/10.1016/j.scitotenv.2020.136729

Ministry of Health. (2005). *Regulation on waters intended for human consumption* (Official Gazette No. 25730, February 17, 2005). Republic of Türkiye Ministry of Health.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). John Wiley & Sons.

Najah, A., El-Shafie, A., Karim, O. A., & El-Shafie, A. H. (2013). Application of artificial neural networks for water quality prediction. *Neural Computing and Applications*, *22*(Suppl 1), 187–201. https://doi.org/10.1007/s00521-012-0940-3

Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, Article 102920. https://doi.org/10.1016/j.jwpe.2022.102920

Panigrahi, N., Patro, S. G. K., Kumar, R., Omar, M., Tran, T. N., Nguyen, L. G., Bui, T. T., & Nguyen, T. T. (2023). Groundwater quality analysis and drinkability prediction using artificial intelligence. *Earth Science Informatics*, *16*(2), 1701–1725. https://doi.org/10.1007/s12145-023-00977-x

Parmar, K. S., & Bhardwaj, R. (2015). Statistical, time series, and fractal analysis of water quality of river Yamuna at Kalpi for forecasting and management. *Environmental Monitoring and Assessment*, *187*(3), Article 124. https://doi.org/10.1007/s10661-015-4338-5

Poudel, D., Shrestha, D., Bhattarai, S., & Ghimire, A. (2023). Comparison of machine learning algorithms in statistically imputed water potability dataset. *Journal of Innovations in Engineering Education*, *5*(1), 38–46. https://doi.org/10.3126/jiee.v5i1.42265

Sathyanarayanan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research*, *27*(4s), 4023–4031. https://africanjournalofbiomedicalresearch.com/index.php/AJBR

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Sprague, L. A., Hirsch, R. M., & Aulenbach, B. T. (2017). Trends in nutrient and sediment concentrations and loads in major river basins of the United States. *Journal of the American Water Resources Association*, *53*(3), 665–684. https://doi.org/10.1111/1752-1688.12520

Turan, T. (2023). Makine öğrenmesi algoritmaları ile su kalitesi ve içilebilirlik tahmini [Water quality and drinkability prediction using machine learning algorithms]. *Uluborlu Mesleki Bilimler Dergisi*, *6*(2), 65–80.

United Nations. (2024). *The United Nations world water development report 2024: Water for prosperity and peace*. UNESCO. https://www.unwater.org

WHO. (2022). World Health Organization. *Drinking-water*. https://www.who.int/news-room/fact-sheets/detail/drinking-water

WHO. (2023). World Health Organization *Drinking-water*. https://www.who.int/news-room/fact-sheets/detail/drinking-water

Zhu, M., Wang, J., Yang, X., & Zhang, Y. (2019). Forecasting water quality parameters in Lake Taihu using machine learning models. *Water Resources Research*, *55*(11), 9502–9519. https://doi.org/10.1029/2019WR025189