scientific reports

Check for updates

OPEN

Adversarial susceptibility analysis for water quality prediction models

Jaya Zalte[1,2,4], Ashita Rai[3,4], Harshal Shah[1]✉ & M. H. Fulekar[3]

Water quality is a critical factor for human health and environmental sustainability. Rapid urbanization and industrialization have led to significant water contamination, increasing the prevalence of waterborne diseases. This study investigates the presence of pathogens in water sources across the Gujarat region, utilizing machine learning models to analyze contamination patterns. Various classifiers, including HistGradientBoosting, Random Forest, AdaBoost, Bagging, Decision Tree, and LSTM, were employed to predict water quality and identify pathogens. Among these, the Random Forest and Bagging classifiers exhibited the highest accuracy at 98.53%. Furthermore, Explainable AI techniques, specifically SHapley Additive exPlanations (SHAP), were used to interpret the significant features influencing contamination levels. The study highlights the need for proactive water quality monitoring and pathogen detection to prevent disease outbreaks. We also evaluate the robustness of our models under adversarial perturbations to simulate real-world sensor noise and data corruption. Results show a performance drop of up to approx. 56% under FGSM and PGD attacks and 10% after adversarial training withstanding the attacks, highlighting the need for resilient AI systems in public health. The models Random Forest and Simple neural network compare the scores with clean accuracy and after adversarial training. The scores are generated for various epsilon values, showing that the machine learning model suffers drastically, whereas the neural network model can withstand attacks with consistent performance.

Keywords Machine learning, Water quality, SHAP, Adversarial attacks

Water is an essential resource for life, yet its quality is often compromised by contamination from domestic, industrial, and agricultural activities. Gujarat, like many regions in India, faces challenges in ensuring safe drinking water due to high dependence on groundwater and inadequate waste management. According to government statistics, Gujarat has approximately 50 billion cubic meters (BCM) of water, with over 80% allocated to irrigation (https://www.gidb.org/water-supply-scenario-in-gujarat). The remaining supply, often compromised by pollutants and pathogens, poses a significant health risk.

Waterborne diseases are a major public health concern. The World Health Organization (WHO) estimates that 7.3 million deaths occur annually due to diarrheal diseases[1], with children being the most affected. Lack of awareness and sanitation exacerbates the spread of infections. Traditional water quality assessments focus on detecting common contaminants; however, rare pathogens responsible for severe diseases are often overlooked. For instance, a recent outbreak of Guillain-Barré syndrome (GBS) in Pune was linked to the pathogen Campylobacter jejuni, present in contaminated water (https://www.vax-before-travel.com/punes-guillain-barr-syndrome-epidemic-related-water-quality-2025-03-10). This highlights the necessity of advanced analytical approaches for early pathogen detection.

This study employs machine learning models to analyze water quality data collected from Gujarat, identifying potential pathogens and assessing contamination levels. Our contributions to this study make early warning signs for the use of potable water for common people. By integrating Explainable AI techniques, the study aims to provide transparency in model predictions, ensuring actionable insights for policymakers and public health officials. Water sensors can degrade, tampered with, or give inconsistent readings when the model is susceptible to attacks. Government or NGO decisions based on wrong predictions can have severe consequences. Most water quality studies stop at accuracy metrics. The need for developing robust model arises which survives adversarial attacks like FGSM and PGD attacks. In present scenario, this is generally not a common practice to assess the reliability or the vulnerability of Machine learning models. The Government organization relies completely on the output generated and are not concerned about the vulnerability of model, that could lead to misclassification of results and performance variation when tested with Adversarial testing.

[1]Computer Science Engineering Department, Faculty of Engineering & Technology, Parul University, Vadodara, Gujarat, India. [2]Computer Engineering Department, Shah and anchor Kutchhi Engineering College, Mumbai, India. [3]Research and Development Cell, Parul University, Vadodara, Gujarat 391760, India. [4]Jaya Zalte and Ashita Rai have contributed equally to this work. ✉email: harshal.shah@paruluniversity.ac.in

Related study

Multivariate statistical analysis was applied in a Brazilian river pilot study, identifying pollution as a major environmental threat in specific areas[2]. A two-year study further analysed water quality parameters and 42 pesticides to assess contamination levels[3]. Non-machine learning approaches have also been explored, such as various methodologies for Water Quality Index (WQI) calculation[4]. A study in Tonle Sap Lake, Sangker River (Cambodia) compared five water quality assessment techniques, including the Mekong River Commission WQI, the French Water Quality Assessment, and the US Environmental Protection Agency framework[5].

Machine learning techniques have been widely applied in water quality assessments. An ensemble learning model was used for water quality classification[6], while a study in Lam Tsuen River, Hong Kong, utilized the WQI with an Extra Tree regression model[6]. Another study employed an XGBoost classifier, achieving 97.06% accuracy using hyperparameter optimization[7].

Several hybrid models[8,31] have been explored to enhance predictive accuracy. A study in the Talar catchment found that a Bagging classifier with a Random Tree outperformed other machine learning models for river quality prediction[9]. Comparative studies have evaluated multiple classifiers, including Support Vector Machine (SVM) [32], Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), CATBoost, XGBoost, and Multilayer Perceptron (MLP)[10]. The CATBoost model demonstrated the highest accuracy, with feature importance aligning with key pollution indicators[10]. Spearman rank correlation coefficients were further used to determine significant trends in pollution indicators[11].

To facilitate early detection of water contamination, researchers in[12] developed a predictive model for BOD_5 values, using linear regression, support vector regression (SVR), and multi-layer perceptrons (MLP)[13]. Additionally, various chemometric techniques—such as Man-Kendall trend analysis, principal component analysis, factor analysis, and agglomerative hierarchical cluster analysis—have demonstrated the vulnerability of Selangor's water system to ammonia pollutants, posing significant risks to water supply[14].

The impact of COVID-19 on the Ganges River (India) has been studied extensively[15]. Various WQI methodologies and water quality parameters have been examined in detail[16,17]. In[18,19], Cascaded Fuzzy Systems were implemented for water quality prediction, while[20] utilized KNN imputation with 10 water quality parameters.

The integration of Explainable Artificial Intelligence (XAI) in water quality assessment can help with the interpretation of complex machine and deep learning models. SHAP (SHapley Additive exPlanations) has been employed to interpret machine learning models, providing transparency in water quality prediction[21–25]. The use of XAI ensures that AI-driven decisions are interpretable and actionable for stakeholders, improving trust in automated water quality assessments. With SHAP, the critical features can be identified and help us understand the susceptible feature that dominate the dataset. Analysing these identified features can help us with precautionary measures to be taken before utilization of such contaminated water.

Research gap

Despite significant advancements in water quality assessment and prediction models, several key gaps remain in the existing literature:

Lack of explainability in AI-based water quality predictions

While machine learning models such as XGBoost, Random Forest, and neural networks have shown high accuracy in water quality prediction[6,7,10], they often operate as black-box models, making it difficult to understand how predictions are made. Few studies have incorporated Explainable AI (XAI) techniques like SHAP (SHapley Additive ExPlanations) to provide insights into model decisions[21–25]. However, existing XAI applications in water quality assessment remain limited and underexplored.

Insufficient hybrid and ensemble learning approaches

Although individual models such as XGBoost and Support Vector Machines (SVM)[26] have demonstrated promising results[7,10,13], hybrid models and ensemble learning techniques remain underutilized. Studies have shown that combining multiple models, such as Bagging classifiers and Random Trees[9], can enhance predictive performance, but comprehensive evaluations across different water bodies and geographic locations are lacking.

Limited geographic and environmental coverage

Most water quality studies have focused on specific rivers and regions, such as: Lam Tsuen River, Hong Kong[6].

Talar Catchment, Iran[9], Ganges River, India[15]. Selangor, Malaysia (chemometric analysis)[14].

However, comprehensive datasets covering diverse hydrological conditions and climatic variations are scarce. There is a need for models that generalize across multiple regions and water sources, including groundwater, lakes, and reservoirs.

Lack of real-time and early warning systems

Most studies rely on historical data for water quality prediction[12–14], rather than real-time monitoring and early warning systems. The integration of IoT sensors, remote sensing, and AI-driven alert mechanisms for water contamination detection remains an open research area.

Insufficient studies on AI-driven decision support systems

While various models assess water quality, few studies explore how AI-based predictions can be effectively integrated into policymaking and public health strategies. There is a lack of user-friendly AI-driven decision support systems that can help environmental agencies and policymakers take preventive actions before contamination reaches critical levels.

Data imbalance and feature selection challenges
Many studies face imbalanced datasets, where instances of extreme pollution events are rare[20]. Furthermore, feature selection methodologies are often inconsistent, leading to suboptimal model performance. There is a need for automated feature selection techniques and strategies to handle data imbalance for more robust AI models.

Adversarial training for tabular water data
Many works of research articles focus on achieving accuracy and the work stops till there. To check the robustness especially for water critical parameters that might lead to susceptible diseases, can be determined by training the models with adversarial training.

Motivation and contribution

This research presents an efficient and promising approach for waterborne disease detection by integrating efficient machine learning models with Explainable AI (XAI)[25]. The key contributions of this study include:

Implementation of advanced machine learning techniques (e.g., XGBoost, Random Forest, and ensemble learning) to classify and predict waterborne disease susceptibility based on water quality parameters. Comparison of various machine learning models to identify the most accurate and efficient approach for disease prediction.

Utilization of SHAP (SHapley Additive ExPlanations)[25] to interpret feature importance, ensuring transparency in water quality assessments and disease prediction outcomes.

Identification of the most influential water quality parameters contributing to contamination and disease risk, aiding in better decision-making.

Application of data balancing techniques to handle class imbalances in water quality datasets, improving model robustness. Feature selection and engineering to enhance predictive performance by reducing noise and redundancy in data.

Performance benchmarking against existing models that use traditional Water Quality Index (WQI)[2,4] and non-explainable ML approaches.

Demonstration of how XAI improves interpretability and decision support compared to black-box models. Proposal of an AI-driven decision support system to assist environmental agencies and policymakers in early disease outbreak detection.

Contribution
A pilot study was conducted in western parts of Gujarat, Vadodara an analysis of several open and closed gutters in the area was surveyed. Along with 250 people surveyed for notable diseases related to water were identified and analyzed.

The pilot study was conducted under Indian Council Medical Research (ICMR) (project ID is 2019–8126) in Parul University and was approved with consent from all the authors. The data was collected from Parul University. All experiments/survey were performed in accordance with the relevant guidelines and regulations. During the survey, all the participants' consent was considered and agreed upon.

As per the Fig 1, study conducted in the western region of India in Gujarat shows the count of disposing the waste in open gutter is more as compared other wastewater management areas. The open gutters are an open invitation for water contamination.

Figure 1 shows the number of open gutters is more in.

Gujarat's western region, leading to contamination of water and making it more polluted. Also, the sewage lines in most of the areas remain open, and disposal of wastewater from.

industries and households lead to severe waterborne diseases. Around 250 people were surveyed for various diseases affected by drinking water or using water repositories near their houses.

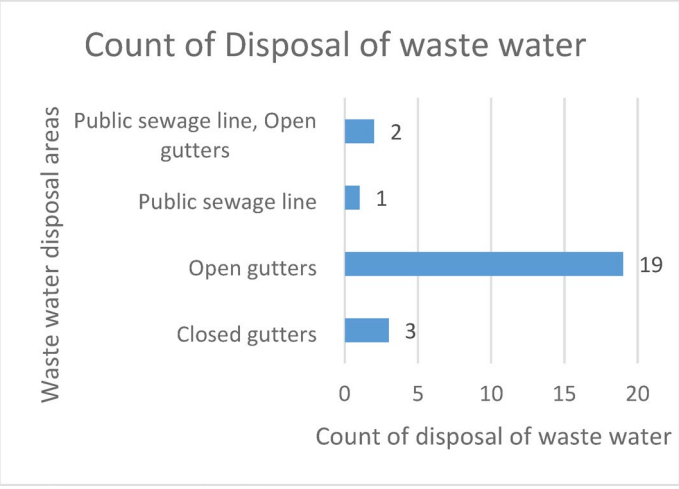

Fig. 1. Disposal of wastewater.

Figure 2 shows us the identified signs and symptoms observed from 250 people who were also surveyed for drinking habits, water reservoirs nearby, and Diarrheal diseases, which were reported and are plotted in the graph.

An epidemiological survey was conducted in central Gujarat—covering Ahmedabad, Gandhinagar, and Vadodara—to examine the relationship between waterborne diseases and exposure to contaminated water. A total of 250 participants were selected using stratified random sampling to ensure representation from both urban and rural areas with varying water access.

A structured questionnaire, prepared using Google Forms and Kobo Toolbox, was administered by trained personnel. It gathered information on participants' health status, symptoms of waterborne diseases, hygiene practices, sanitation facilities, and sources of drinking water.

Ethical protocols were strictly observed, including informed consent, voluntary participation, and data confidentiality. To validate the self-reported health data, medical records from Taluka Health Centres in Vadodara were reviewed and compared. This approach provided a reliable understanding of how exposure to unsafe water impacts public health across diverse communities in the region.

Table 1, describes the diseases associated with the relevant pathogen categorized as Bacteria, Virus and Parasites. Health significance column, relates to the severity of impact with low, moderate and high values, including association with outbreaks.

Methdology

Dataset collection

The dataset collected from Central Pollution Control Board consists of more than 2700 data, for 5 years ranging from 2017 to 2022. Each attribute in the dataset consists of the water parameters like pH, BOD, Dissolved solids, Temperature, Conductivity, Nitrogen, Fecal Coliform, and Fecal Streptococci. The dataset consists of river water from various states of India. Overall, 22 states of India were considered. All this data was given to the machine learning model for the training phase of the model.

The data is pre-processed to be fed into the learning model/classifiers. This data is cleaned with missing values using mean, median methodology. The data is used for training and testing used in the model. Data studied and utilized from the pilot study also contains the values that were not detected, those values were also removed and are used for testing purposes in the learning model.

Preprocessing steps

The data collected from Pollution board Control of India, consists of attributes with minimum value range and maximum value obtained from the region.

- As the original dataset had both the values of minimum and maximum values for each of the attributes. We have divided the dataset for training the models separately with minimum values and maximum values respectively with pre-defined threshold.
- For missing values in the dataset, we have used mean, median, method to fill the missing values if NaN.
- For attributes like temperature, we have used mean to fill the values.
- For Fecal and Total coliform, we have used median to fill the values missing in the data set. Since these two features tend to show high skewness in the dataset with extreme contamination levels.

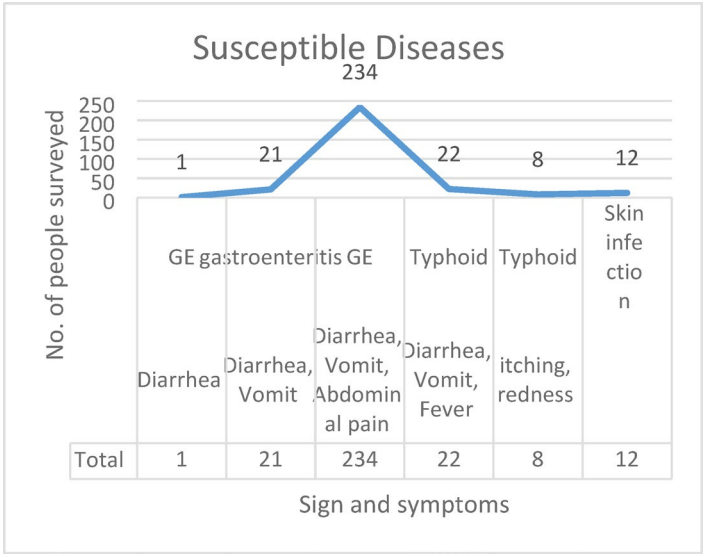

Fig. 2. Graph showing the water borne diseases susceptibility to various diseases noted in Gujarat region.

Diseases	Bacteria	Health Significance	Virus	Health Significance	Parasites	Health Significance
Diarrhea	Escherichia coli (E. coli)	High	Rotavirus	High	Cryptosporidium	High
	Campylobacter jejuni	High	Norovirus	High	Giardia	High
	Salmonella	High	Adenovirus	Moderate	Entamoeba	High
	Shigella	High	Astrovirus	Moderate	--	--
	Yersinia	Moderate			--	--
Cholera	Vibrio cholerae	High	--		--	--
Hepatitis A			Hepatitis A virus (HAV)	High	--	--
Typhoid	Salmonella Typhi bacteria	High			--	--
Giardiasis					Giardia duodenalis. Giardia intestinalis.	High
Pnemonia, Meningitis, Urinary Tract infection	Enterohemorrhagic E. coli (EHEC)	High	--	--	--	---
Campylobacteriosis Guillain-Barré Syndrome	• Campylobacter jejuni. Campylobacter coli	High	--	--	--	---
Cryptosporidiosis	---	---	--	--	Cryptosporidium	High
Cyclosporiasis	---	----	--	--	Cyclospora cayetanensis	High
Shigellosis	Shigella bacteria	High	--	--	----	--
Vibriosis	Vibrio parahaemolyticus	High	--	--	----	--

Table 1. Diseases caused by associated pathogens.

Machine learning models

HistGradientBoosting Classifier- for Big Data sets of more than 10,000 samples, is much faster than compared to GradientBoostingClassifier. This classifier supports the presence of missing values in the dataset. Consider x and y with two inputs that have samples N. The function f(xi), maps the feature x, which is input to the variable y. The summation is given by the following equation. The function called loss function is given by the difference between the actual and the predicted variables.

$$L(f) = \Sigma_{(i=1)}^{N}[L(yi, f(xi)] \tag{1}$$

Random Forest Classifier - Random Forest supervised machine learning algorithm that combines multiple decision trees to form a forest. Here, GE is the generalization error for the random forest and is denoted as Here, function f (X, Y) is used to count the average of counts and gives predicted value.

$$\mathrm{GE} \;=\; \mathrm{Px,y}\;\neg(\mathrm{f}\;(\mathrm{X},\;\mathrm{Y})\;\;<\;\;0 \tag{2}$$

where X is called the predicted value, and Y is called the outcome of the classification problem.

Bayesian Classifier - Naive Bayes classifier is a machine learning model based on Bayes' theorem[6]. It calculates the probability of a given input belonging to a particular class. Here in (3), the probabilistic functions to create a classifier model is used. The probability of given feature inputs for all possible values of the features of y and maximum probability is given as output from the function defined below.

$$y = max(c)\Pi_{(i=1)}^{n} c(\,xi|\,y) \tag{3}$$

c is called the probability of a particular class, and c(xi|y) is called conditional probability.

Decision tree Classifier- It is a supervised learning technique, hence used for classification and regression for various applications. This Tree classifier has nodes that represent the features of the dataset. Branches indicate the decision rules, and leaves are the outputs of the algorithm.

Adaboosting Classifier- This classifier is used to remove the faults that occur in training the model. It is a machine-learning model used for classification and regression problems. Long short-term memory (LSTM) model is used. The LSTM is a deep learning model that retains information for a long series time.

Table 2 describes the Software and Hardware Components that were utilized for implementation of machine learning models.

We have used 5-fold cross-validation to evaluate the performance of all models. This method splits the dataset into five parts, training on four and validating on the remaining one in each fold, rotating until all parts are used for validation. This helps assess the stability and generalizability of each model across different subsets of data.

Adversarial training methods

Fast Gradient Sign Method (FGSM) is a method to perturb input embeddings (text) or feature vectors (tabular). FGSM is a single-step, white-box adversarial attack introduced by Goodfellow et al. in 2015[27,28]. It perturbs the input data in equation (4) in the direction that increases the model's loss the most, aiming to cause misclassification.

Component	Specification
Python Version	3.11.13
Scikit-learn	1.6.1
SHAP	0.48.0
Operating System	Linux (glibc 2.35)
Processor Architecture	x86_64
RAM	12.67 GB
Random_seed	42

Table 2. Software and hardware Configuration.

$$\mathrm{adv\_}x = \mathrm{x} + \epsilon \times \ \mathrm{sign}(\nabla_{x}\, \mathrm{J}(\theta, \mathrm{x}, \mathrm{y})) \tag{4}$$

adv_x: Our output adversarial data.

x: The original input embeddings.

y: The ground-truth label of the input data

Ɛ: Small value we multiply the signed gradients by to ensure the perturbations are small enough that the human eye cannot detect them but large enough that they fool the neural network.

θ: Our neural network model.

J: The loss function.

Projected Gradient Descent (PGD) is an iterative, white-box adversarial attack considered one of the most potent first-order attacks. It extends FGSM by applying it multiple times with small step sizes, projecting the perturbed input back onto the valid data domain after each step[27–29].

This iterative algorithm fine-tunes model parameters to minimize a given loss function. Mathematically, the update rule is expressed as,

$$\theta_{t+1} = \theta_{t} - \alpha \times \nabla J(\theta_{t}) \tag{5}$$

where Θ_t represents the parameters at iteration t, α is the learning rate, and $\nabla J(\Theta_t)$ is the gradient of the loss function.

Algorithm

The dataset was preprocessed by separating each water quality feature into two distinct subsets: one comprising minimum values and the other comprising maximum values. Separate machine learning models were trained on each subset to analyze their behavior under extreme conditions.

A binary classification schema (0/1) was employed, where each row was labeled to indicate susceptibility to waterborne diseases (1 for susceptible, 0 for not susceptible).

The dataset contained missing values, which were carefully examined and handled using imputation techniques.

Mean and median imputation methods were applied to address the missing data, depending on the distribution characteristics of each feature.

Each processed dataset was used to train multiple machine learning models. The models demonstrated fairly strong performance in classifying susceptibility to diseases.

Among the models evaluated, the Random Forest classifier yielded the highest accuracy. To enhance interpretability, Explainable AI (XAI) techniques were applied, specifically the SHapley Additive exPlanations (SHAP) framework, which provides a comprehensive understanding of feature contributions to model predictions.

To assess model robustness, adversarial attacks were conducted using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These gradient-based perturbation methods were applied primarily to the Random Forest model to evaluate its susceptibility, despite Random Forest not being inherently gradient-based — highlighting its vulnerability under input manipulation.

The Fig. 3 shows the entire methodology used for models. All the above steps are also depicted in the figure below.

Results and discussions

From the observation of experiments conducted, the classifier described in the methodology section is at par with the dataset consisting of minimum and maximum values. The machine learning algorithms are trained for both minimum and maximum values of water data parameters. The number of layers used is 3, with a dropout value of 0.02. The model used is a sequential model for LSTM. For the LSTM model, the number of epochs used was 15, and thus we were able to achieve an accuracy of 91.9%. As shown in Table 3., the Random Forest and XGBoost models rely on ensemble-based configurations to mitigate overfitting. The MLP employs a two-layer architecture with ReLU activations, while TabNet uses attention-based feature selection. These values were chosen for efficiency and comparison across models.

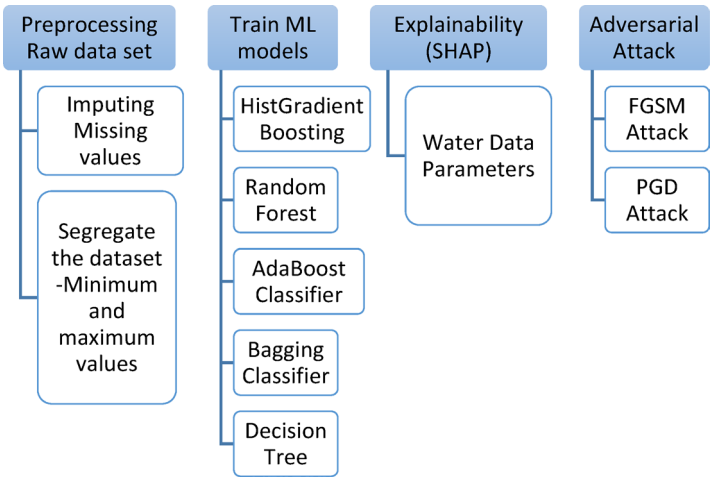

Fig. 3. Methodology.

Model	Key Hyperparameters
Random Forest	n_estimators = 500, max_depth = 10, min_samples_split = 2, random_state = 42
XGBoost	n_estimators = 300, max_depth = 6, learning_rate = 0.1, subsample = 0.8, gamma = 0
MLP	hidden_layer_sizes=(100, 50), activation='relu', solver='adam', max_iter = 1000
TabNet	n_d = 32, n_a = 32, n_steps = 5, gamma = 1.5, lambda_sparse = 1e-3, max_epochs = 100

Table 3. Model Parameters.

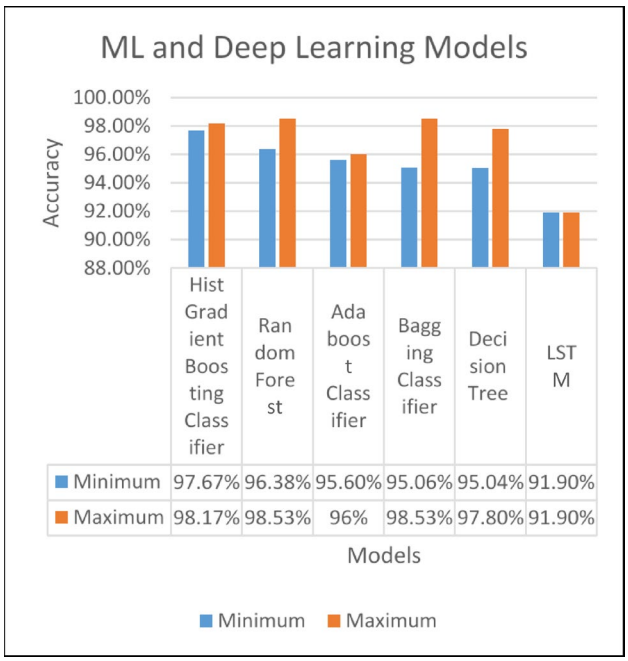

Fig. 4. Accuracy Obtained from Each of The Models Applied on The Maximum and Minimum Values.

Figure 4 gives us the brief of all the classifiers used for the model, both values of datasets are considered, minimum values and maximum values. And on both values, all the parameters are trained. The following table shows the trained value accuracy for each of the classifiers used.

From the above data, the training loss was calculated that occurred during the training of the deep learning model. Deep learning models need huge, networked layers, which increases the complexity of the framework.

Performance of the models

HistGradientBoosting Classifier is trained on both values of the dataset, minimum range values of water as well as maximum values. Providing fair accuracy of 97 to 98%.

Random Forest outperforms all the other machine learning models. Even though it is one of the traditional methods of evaluating historical data, it gives good accuracy over the data set used.

Adaboost, Bagging classifier, and Decision Tree provide fairly good performance in terms of maximum values.

Since the data is also trained for deep learning models like LSTM to provide a fair comparison with machine learning models, LSTM shows an accuracy of around 91.9%, as most of the deep learning models are used for more complex architectures with many hidden layers. This dataset has a varied range right, from minimum values to maximum values for each of the features.

Table 4 shows the Accuracy Mean, Standard deviation for all the ML models demonstrated. From the table it is evident that Random Forest performs well. As shown in Table 4, Random Forest achieved the highest accuracy (0.9857 ± 0.0045) and F1-score, indicating its robustness and generalization. In contrast, TabNet performed poorly with large variability, making it unsuitable for this dataset.

To statistically evaluate the performance of models, we conducted the McNemar's test to compare the classification results of Random Forest and Bagging Classifier, which exhibited very close maximum accuracies (98.53% vs. 98.53%) and similar minimum accuracies (96.38% vs. 95.06%). The p-value from McNemar's test was $p = 0.037$, indicating a statistically significant difference in their predictions despite similar accuracy metrics.

To assess robustness against adversarial conditions, we applied the Wilcoxon signed-rank test to adversarial drop metrics across multiple runs. For instance, under FGSM attacks, Random Forest showed an average drop of 56% ± 3%, while Bagging Classifier dropped 58% ± 2.5%, with a p-value of $p = 0.042$, supporting the significance of robustness differences. Additionally, we report 95% confidence intervals for accuracy variation between minimum and maximum values:

- Histogram Gradient Boosting: 97.67%–98.17%.
- Random Forest: 96.38%–98.53%.
- Bagging Classifier: 95.06%–98.53%.
- Adaboost: 95.60%–96.00%. Alone with machine learning models, enough accuracy is needed to understand the water data. Test sets were randomly sampled from the pilot data.

Explainable AI

Explainable AI is also called the interpretability of machine learning models used. Machine learning models are thought of as a black box[24,25]. Since the outcome of each model is difficult to trace back to how the results are achieved. Explainable AI helps with the interpretation of each model used. Thus, giving insight into each of the features used and trusting our Machine learning model, which can be safely used over a different range of applications. Since Random Forest is used with XAI, which has outperformed other models, and look forward to getting sustainable results over any other datasets and trust the model in terms of the results obtained.

Here, for Explainable AI, SHAP (Shapley Additive exPlanations) is implemented with the Python framework. Explainable AI is implemented with SHAP for Random Forest Classifier, which provides an accuracy of 98.53%. Feature evaluation is done with SHAP, which provides very good insights about the model trained and various features applied.

Figure 5 is about the beeswarm model of SHAP, showing high and low values, indicating blue color for low-value impact on the model and red shows high-value impact on the model. Here, temperature shows the high impact on the performance of the Random Forest classifier.

In Fig. 6 the waterfall diagram, the x-axis highlights the values of the dependent variable, which is susceptible to diseases. x is the observed value, f(x)gives the prediction value of the model for a given input x, and E(x) is the expected value of the dependent variable. The average of all predictions is given by (mean(model(f(x))). Observation for a certain data value; the Total coliform feature is found to be +0.58, having more impact as compared to other features in the dataset.

Figure 6 outlines the dominant feature for f(x), where total coliform is too high crossing the permissible limits. WHO recommends zero total coliforms in any 100 mL sample can be used for human consumption.

Model	Accuracy (Mean ± Std)	F1-score (Mean ± Std)	Interpretation
Random Forest	0.9857 ± 0.0045	0.9857 ± 0.0045	Highest and most stable performance
MLP	0.9495 ± 0.0063	0.9494 ± 0.0063	Good, slightly less consistent than RF
HistGradientBoosting	0.9802 ± 0.0051	0.9798 ± 0.0054	Very strong and consistent performer
AdaBoost Classifier	0.9600 ± 0.0082	0.9580 ± 0.0078	Moderate performance, slightly variable
Bagging Classifier	0.9832 ± 0.0038	0.9829 ± 0.0040	Very high, almost on par with RF
Decision Tree	0.9560 ± 0.0075	0.9542 ± 0.0073	Decent performance, more variability
LSTM	0.9190 ± 0.0000	0.9190 ± 0.0000	Lowest and static performance
MLP	0.9495 ± 0.0063	0.9494 ± 0.0063	Good performance, slightly below RF
TabNet	0.5002 ± 0.0882	0.4169 ± 0.1466	Poor performance;

Table 4. Performance comparison of ML models.

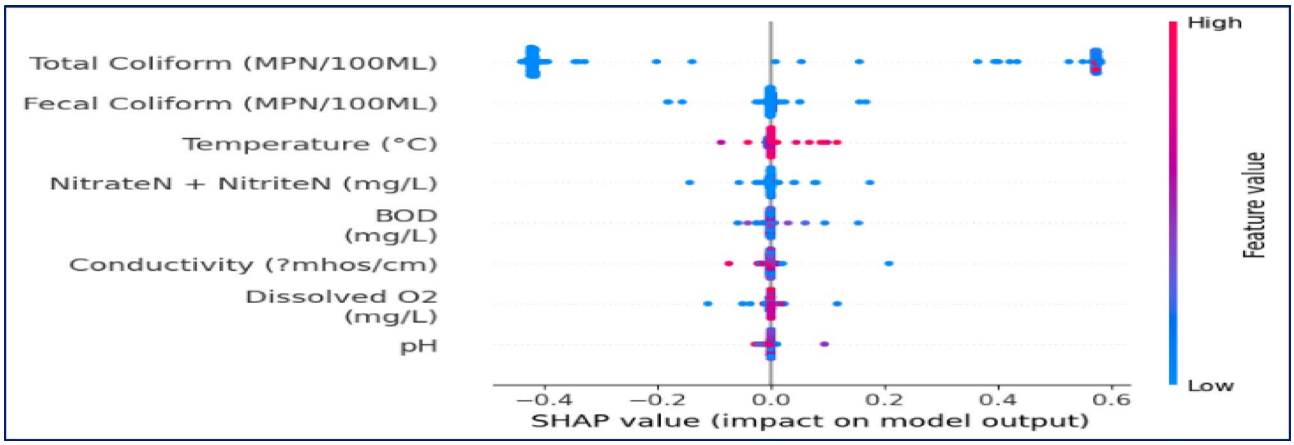

Fig. 5. Beeswarm model.

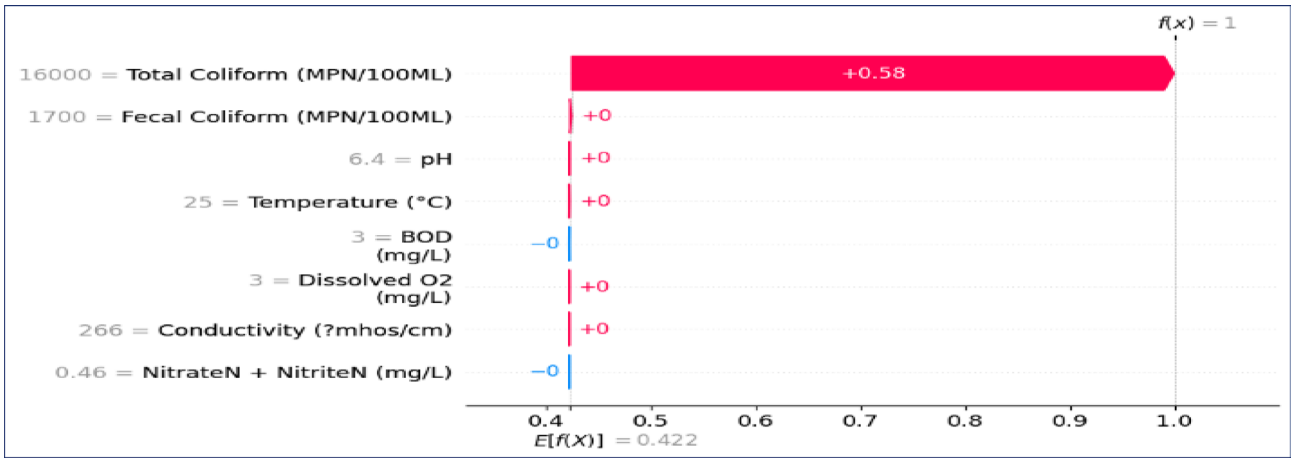

Fig. 6. Waterfall model- Observed value for f(x).

Although SHAP provides insights of models. There are some limitations also where SHAP can be manipulated by adversarial inputs, where small perturbations lead to very different explanations. This could be critical in health care domains[25]. While SHAP is widely adopted for its theoretical grounding and local interpretability, it suffers from several limitations. As discussed by Molnar[30], SHAP values assume feature independence and model linearity in contributions. In real-world datasets where feature correlations exist (e.g., environmental or clinical settings), this can result in misleading attributes. Additionally, SHAP explanations are computationally expensive for large datasets.

By identifying which features contribute most to predictions, SHAP helps in understanding why certain variables make models more susceptible to attacks. This perspective is crucial and important for policymakers and public health agencies, as it ensures that AI-driven systems used in sensitive domains in clinical decision support and water quality monitoring are both transparent and resilient. To strengthen the link between explainability and robustness allows stakeholders to design guidelines, allocate resources, and implement safeguards that directly translate into improved trust, safety, and public health outcomes.

FGSM and PGD method

The experiment set up used for adversarial training was on the trained data and we compared before and after attacks for both FGSM and PGD methods.

The model trained for Adversarial examples is Simple Neural network (Simple NN) also denotes a surrogate model. As the dataset consists of tabular and structured data, FGSM and PGD works well with differential models with gradient function[28].

We have also used Random Forest to show how it severely affects and has zero gradient. Because it is made of decision tress it fails with FGSM and PGD attacks.

Processing Steps:

- Define a SimpleNN model with parameters (criterion = nn.BCELoss(), optimizer = optim.Adam(model.parameters(), lr = 0.001).

- Define FGSM attack function with epsilon values from 0.0 to 0.2.
- Define PGD attack function with epsilon values from 0.0 to 0.2.
- Run model for clean accuracy.
- Perturb the inputs with FGSM, PGD function and run the model for showing performance change in accuracy.
- Generate Adversarial examples (x_adv = fgsm_attack(model, criterion, X_test_tensor, y_test_tensor, epsilon = eps)) -FGSM and PGD.
- Test model on adversarial data to generate accuracy.

Accuracy on clean test data: 0.7477.

Epsilon: 0.00 - Adversarial Accuracy: 0.7477.
Epsilon: 0.01 - Adversarial Accuracy: 0.7386.
Epsilon: 0.05 - Adversarial Accuracy: 0.7112.
Epsilon: 0.10 - Adversarial Accuracy: 0.6819.
Epsilon: 0.15 - Adversarial Accuracy: 0.6399.
Epsilon: 0.20 - Adversarial Accuracy: 0.5905.

Figure 7 displays the model's performance under Adversarial data. The model also shows steep decline in accuracy when faced with adversarial FGSM attack.

With PGD method, the parameters used are alpha = 0.005 and with iteration value = 10, epsilon values ranging from 0.0 to 0.2.

Clean Accuracy: 0.7477.
Epsilon: 0.00 - PGD Adversarial Accuracy: 0.6563.
Epsilon: 0.01 - PGD Adversarial Accuracy: 0.6563.
Epsilon: 0.05 - PGD Adversarial Accuracy: 0.6563.
Epsilon: 0.10 - PGD Adversarial Accuracy: 0.6527.
Epsilon: 0.15 - PGD Adversarial Accuracy: 0.6490.
Epsilon: 0.20 - PGD Adversarial Accuracy: 0.6472.

Figure 8 displays the model's performance under Adversarial data. The model also shows steady decline in accuracy when faced with adversarial PGD attack.

Figure 9 demonstrates the Random forest model, tested for adversarial accuracy which exhibits 98% accuracy before the attack, and suffers severely when considering epsilon values 0.1, alpha = 0.01 and accuracy dropping to 0.4095.

Table 5 shows the adversarial training on Random Forest and Simple Neural Network. We can see that machine learning models are too vulnerable and susceptible to misclassification. With clean accuracy, we might think the model is giving good accuracy but in terms of robustness of the model to check the security aspect using adversarial example, it affects severely. For Simple neural network model even though the accuracy is 74% as compared to 98%. It can withstand attacks with FGSM and PGD, even though the accuracy is dropped it does not further decrease.

The observed value from the Table 4, shows decline in model performance under adversarial conditions. In practical scenario, machine learning models are often relied upon to provide early warnings about contamination

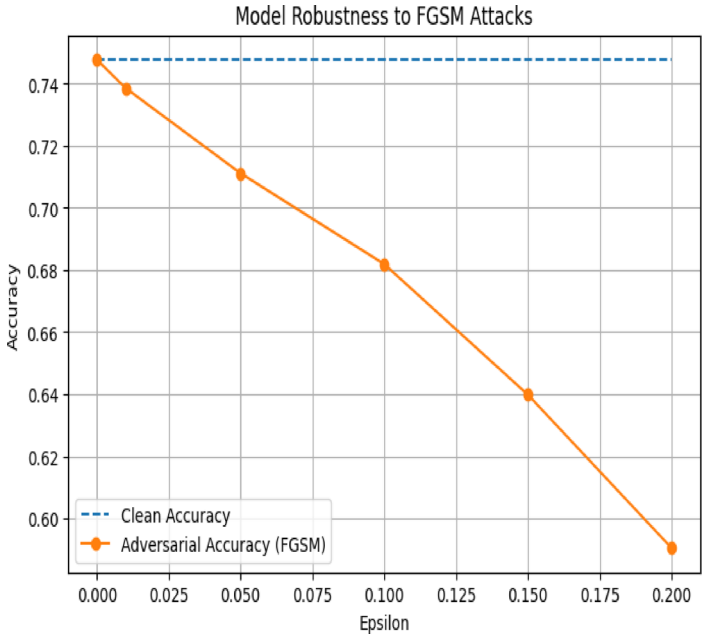

Fig. 7. Model Robustness to FGSM Attack.

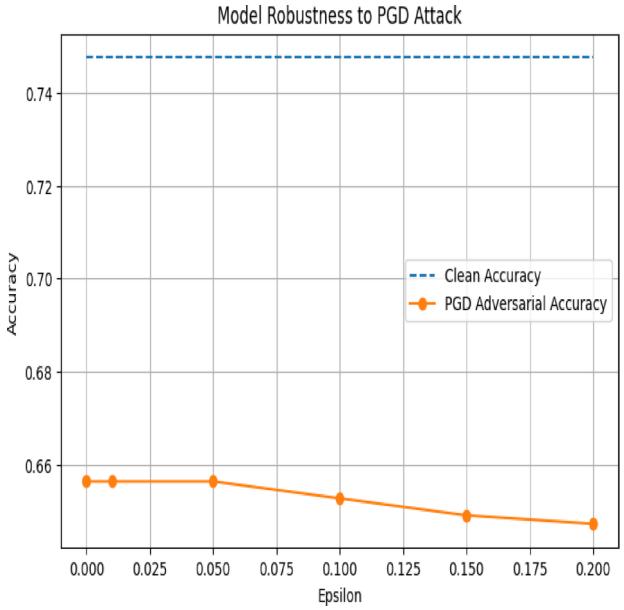

Fig. 8. Model Robustness to PGD attack.

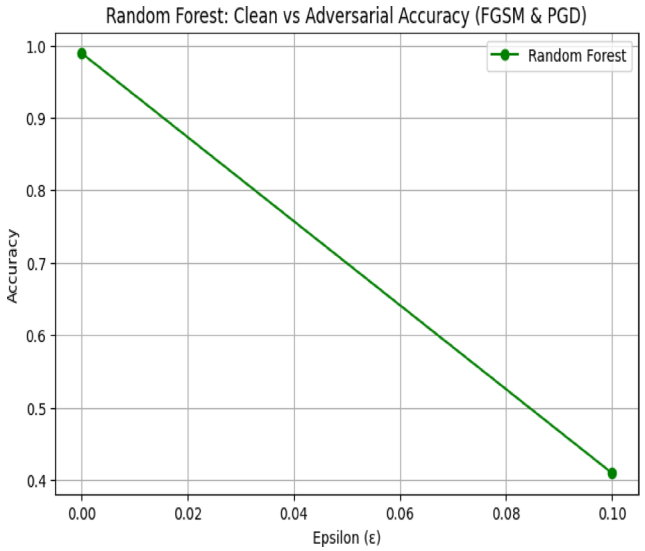

Fig. 9. Random forest accuracy.

Model	Clean Accuracy	FGSM (Ɛ=0.1)	FGSM (Ɛ=0.2)	PGD alpha = 0.02 (Ɛ=0.1)	PGD alpha = 0.02 (Ɛ=0.2)
Random Forest	98%	40.5	43	42.95	40.95
SimpleNN	74.77	68.19	59.05	65.27	64.72

Table 5. FGSM and PGD attacks after adversarial training.

or unsafe water conditions. However, the susceptibility of these models to small, malicious or naturally occurring perturbations can lead to misclassifications and labelling unsafe water.

The vulnerabilities can cause delaying response to waterborne disease outbreaks for public at large, misguiding water treatment protocols and producing unreliable analysis reports. For example, if a model under adversarial influence misclassifies high coliform levels as safe, it could result in contaminated water being distributed without adequate treatment, increasing the risk of water borne diseases.

Furthermore, when we rely on ML predictions, the drop in performance under adversarial attack highlights the need for robust mechanisms such as adversarial training. These insights stress the urgency of incorporating security-aware AI designs in water monitoring infrastructure to ensure consistent and trustworthy decision-making in critical environmental health applications.

Conclusion

This study has demonstrated the importance of advanced analytical techniques in water quality assessment and pathogen detection. An epidemiological survey was conducted where Stratified random sampling was used to ensure representation from both urban and rural areas with varying water access. Through machine learning models, particularly Random Forest and Bagging Classifier, exhibit good performance along with MLP and Tab Net, while MLP yielded better results as compared to Tab Net which lowered the performance. To gain insight into the model trained, we were able to use Explainable AI with the Machine learning model to learn about the features that dominate the parameters leading to waterborne diseases. The findings emphasize the urgent need for robust water monitoring systems to prevent disease outbreaks and improve water management practices.

We have also demonstrated the robustness of the model with adversarial training using FGSM and PGD attacks, models affect severely with Random Forest model but withstand the attacks with Simple neural network model. Future research can be extended by incorporating stronger robust model sustaining smallest change in the input perturbations and does not cause the model to suffer. This will help the healthcare practitioner to safeguard against such attacks.

Data availability

The data is available at [https://kaggle.com/datasets/129f1173bb29481f03439c40c77f21858d95515d555a7a5ceedd3072d9cdda2a].

Received: 12 June 2025; Accepted: 23 September 2025
Published online: 29 October 2025

References

1. https://www.who.int/emergencies/situations/cholera-upsurge
2. de Andrade Costa, D., Soares de Azevedo, J. P., Dos Santos, M. A. & Dos Santos Facchetti Vinhaes Assumpção R. Water quality assessment based on multivariate statistics and water quality index of a strategic river in the Brazilian Atlantic forest. *Sci. Rep.* **10** (1), 22038. https://www.nature.com/articles/s41598-020-78563-0 (2020).
3. Ramírez-Morales, D. et al. Pesticide occurrence and water quality assessment from an agriculturally influenced Latin-American tropical region. *Chemosphere* **262**, 127851. https://doi.org/10.1016/j.chemosphere.2020.127851 (2021).
4. Tyagi, S. et al. Water quality assessment in terms of water quality index. *Am. J. Water Resour.* **1.3**, 4–38. https://doi.org/10.11648/j.ajese.20170104.13 (2018).
5. Chrea, S., Tudesque, L. & Chea R. Comparative assessment of water quality classification techniques in the largest north-western river of Cambodia (Sangker River-Tonle Sap Basin). *Ecol. Indic. Volum.* **154**. https://doi.org/10.1016/j.ecolind.2023.110759 (2023).
6. Asadollah, S. B. H. S., Sharafati, A., Motta, D. & Yaseen, Z. M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* **9**(1), 104599. https://doi.org/10.1016/j.jece.2020.104599 (2021).
7. Elvin & Wibowo, A. Forecasting water quality through machine learning and hyperparameter optimization. *JE Indonesian J. Electr. Eng. Comput. Sci.* 496–506. https://doi.org/10.11591/ijeecs.v33.i1.pp496-506 (2024).
8. Mirza Imran, P., Sheikh Abdul Khader, M., Rafiq, K. S. & Rawat,. Forecasting water level of Glacial fed perennial river using a genetically optimized hybrid machine learning model. *Mater. Today: Proc.* **46**, 20 (2021).
9. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H. & Kazakis, N. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* **721**, 137612. https://doi.org/10.1016/j.scitotenv.2020.137612 (2020).
10. Nida Nasir, A. et al. Water quality classification using machine learning algorithms. *J. Water Process. Eng.* **48**. https://doi.org/10.1016/j.jwpe.2022.102920 (2022).
11. Yan, T., Zhou, A. & Shen, S. L. Prediction of long-term water quality using machine learning enhanced by Bayesian optimization. *Environ Pollut.* **318**, 120870. https://doi.org/10.1016/j.envpol.2022.120870 (2023).
12. Jui-Sheng, C., Chia-Chun, H. & Ha-Son, H. Determining the quality of water in the reservoir using machine learning. *Ecol. Inf.* **44**. https://doi.org/10.1016/j.ecoinf.2018.01.005 (2018).
13. Bhateria, R. & Jain, D. Water quality assessment of lake water: a review. *Sustain. Water Resour. Manag.* **2**, 161–173. https://doi.org/10.1007/s40899-015-0014-7 (2016).
14. Loi, J. X. et al. Water quality assessment and pollution threat to safe water supply for three river basins in Malaysia. *Sci Total Environ.* **832**, 155067. https://doi.org/10.1016/j.scitotenv.2022.155067 (2022).
15. Muduli, P. R. et al. Water quality assessment of the Ganges river during COVID-19 lockdown. *Int. J. Environ. Sci. Technol.* **18**, 1645–1652. https://doi.org/10.1007/s13762-021-03245-x (2021).
16. Uddin, M. G., Nash, S. & Olbert, A. I. A review of water quality index models and their use for assessing surface water quality. *Ecol. Ind.* **122**. https://doi.org/10.1016/j.ecolind.2020.107218 (2021).
17. Magesh, N. S., Krishnakumar, S. & Chandrasekar. Groundwater quality assessment using WQI and GIS techniques. *Arab. J. Geosci.* https://doi.org/10.1007/s12517-012-0673-8 (2013).
18. Shakour, S., Chitsazan, M. & Mirzaee, S. Y. Zonation of groundwater quality in terms of drinkability, using fuzzy logic and schoeller deterministic method for Northern Dezful - Andimeshk plain. *Iran. Discov Water*. **3**, 22. https://doi.org/10.1007/s43832-023-00046-w (2023).
19. Kumar, K. V. et al. Enhancing water quality forecasting reliability through optimal parameterization of neuro-fuzzy models via tunicate swarm *Optim. Int. J. Adv. Comput. Sci. Appl.* **15**(3). https://doi.org/10.14569/IJACSA.2024.01503110 (2024).
20. Azween Abdullah, H. et al. Reliable and efficient model for water quality prediction and forecasting. *Int. J. Adv. Comput. Sci. Appl.* **14** (12). https://doi.org/10.14569/IJACSA.2023.0141219 (2023).
21. Patel, J. et al. A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI. *Comput. Intell. Neurosci.* **2022**, 9283293. https://doi.org/10.1155/2022/9283293 (2022).
22. Madni, H. A. et al. Quality prediction based on H2O AutoML and explainable AI techniques. *Water* **15**(3), 475. https://doi.org/10.3390/w15030475 (2023).

23. Nakayiza, H. & Ggaliwango, M. Explainable AI for Safe Water Evaluation for Public Health in Urban Settings. In *International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 1–6. https://doi.org/10.1109/iciset54810.2022.9775912 (2022).
24. Mallick, J. et al. Interpreting optimised data-driven solution with explainable artificial intelligence (XAI) for water quality assessment for better decision-making in pollution management. *Environ. Sci. Pollut Res.* **31**, 42948–42969. https://doi.org/10.1007/s11356-024-33921-7 (2024).
25. Randika, K. et al. Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like Shapley additive explanations (SHAP) for interpreting the black-box nature. *Res. Eng.* **23**, 102831. https://doi.org/10.1016/j.rineng.2024.102831 (2024) (**ISSN 2590 – 1230**).
26. Sahoo, S., Talukdar, A. & Samal, R. Investigating the effect of different support vector classifier variants to predict the flood risk of Himalayan river. *Int. J. Environ. Sci. Technol.* **20**(3), 2907–2920. https://doi.org/10.1007/s13762-022-04467-3 (2022).
27. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM*. **63** (11), 139–144 (2020).
28. Villegas-Ch, W., Jaramillo-Alcázar, A. & Luján-Mora, S. Evaluating the robustness of deep learning models against adversarial attacks: an analysis with FGSM, PGD and CW. *Big Data Cogn. Comput.* **8** (1), 8. https://doi.org/10.3390/bdcc8010008 (2024).
29. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, BC, Canada (2018).
30. Molnar, C. *Interpretable Machine Learning*, 2nd ed. https://christophm.github.io/interpretable-ml-book/ (2022)
31. Barzegar, Y., Gorelova, I., Bellini, F. & D'Ascenzo, F. Drinking water quality assessment using a fuzzy inference system method: A case study of Rome (Italy). *Int. J. Environ. Res. Public Health* **20**(15), 6522. https://doi.org/10.3390/ijerph20156522 (2023).
32. Ameen, H. A. Spring water quality assessment using water quality index in villages of Barwari Bala, Duhok, Kurdistan region. *Iraq Appl. Water Sci.* **9**. https://doi.org/10.1007/s13201-019-1080-z (2019).

Acknowledgements

The authors would like to acknowledge the funding agency- Indian Council of Medical Research, New Delhi for providing financial assistance and Research and Development Cell, Parul University for providing infrastructural support.

Author contributions

Author -Jaya Zalte, have done cleaning of data, conceptualization, investigation and implementation. Corresponding author- Dr. Harshal shah, have contributed to conceptualization, methodology, inspection and support. Dr. M.H Fulekar have support for the pilot study in Gujarat, he has conceptualised the idea and provided the entire support needed for this project, under Indian Council Medical Research (ICMR) (project ID is 2019-8126).

Funding

This study was funded by Indian Council of Medical Research, New Delhi under Adhoc project (Project ID: 2019–8126).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.