Water Potability Classification: A Machine Learning Approach

Harsh Vishwakarma
Roll No. 2022205
harsh22205@iiitd.ac.in

Ayan Kumar Singh
Roll No. 2022122
ayan22122@iiitd.ac.in

November 17, 2025

1 Introduction

Access to safe drinking water is a fundamental human right and a critical global challenge. Water quality is increasingly threatened by a complex array of anthropogenic contaminants, ranging from traditional physicochemical parameters to emerging threats such as Per- and Polyfluoroalkyl Substances (PFAS), industrial effluents, and microbial pathogens. Addressing these challenges requires a dual approach: understanding the behavior and risks of these contaminants through rigorous environmental assessment, and developing robust, automated systems to predict water potability based on available data.

This research presents a comprehensive study that bridges the gap between environmental toxicology and computational intelligence. First, it provides an integrated assessment of water quality deterioration, synthesizing recent findings on contaminant mobility, environmental risk, and mitigation strategies. This theoretical framework highlights the urgency of monitoring complex water matrices, from municipal supplies to treated industrial wastewater.

Second, leveraging this understanding of water chemistry, the study implements a practical machine learning framework to classify water potability. Using the Water Potability dataset, which contains essential physicochemical properties—such as pH, hardness, solids, chloramines, conductivity, and turbidity—this project develops predictive models to determine whether water

is safe for human consumption. The computational workflow involves extensive data preprocessing, including missing value imputation and Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance.

Multiple classification algorithms, including Support Vector Machines (SVM), Random Forest, XGBoost, and LightGBM, are trained and evaluated to identify the most accurate approach. By combining a theoretical understanding of contaminant risks with advanced data-driven classification, this work aims to provide a holistic perspective on water quality assessment, offering both scientific insights and practical tools for decision-making.

1.1 Objectives

The primary objectives of this research are to:

- **Synthesize Current Knowledge:** Conduct an integrated literature review on emerging water contaminants (PFAS, pathogens) and modern mitigation strategies to contextualize the importance of water quality monitoring.
- **Data Processing & Analysis:** Perform rigorous exploratory data analysis (EDA) and preprocessing on the Water Potability dataset, addressing missing values and class imbalance using SMOTE.
- **Model Development:** Develop and compare a diverse range of machine learning classifiers (including Ensemble methods and SVMs) to predict water safety.
- **Optimization:** Enhance model performance through hyperparameter tuning (GridSearchCV) to maximize metrics such as ROC-AUC and F1-Score.
- **Deployment:** Propose a framework for a web-based interface to allow real-time predictions based on the best-performing model.

2 Integrated Assessment of Water Quality Deterioration: Contaminant Behavior, Risk, and Mitigation Strategies

2.1 Introduction to Research Context

Water quality deterioration represents a critical global challenge driven by an expanding array of anthropogenic contaminants. These range from persistent organic pollutants (POPs) like PFAS to microbial pathogens and industrial byproducts. The complexity of modern pollution necessitates a shift from monitoring single contaminants to understanding the cumulative risks posed by complex mixtures in aquatic environments.

While regulatory frameworks have historically focused on well-known pathogens and basic chemical parameters, emerging research highlights several urgent issues:

- The ubiquity of "forever chemicals" (PFAS).
- Underreported viral loads in wastewater.
- The toxicity of industrial effluents like produced water.

The following analysis integrates findings from five recent studies to construct a comprehensive picture of current water quality threats. These studies collectively underscore the inadequacy of current monitoring paradigms and the necessity for advanced treatment and surveillance strategies:

1. **PFAS Occurrence:** Analysis of Canadian drinking water sources [1].
2. **Treatment Co-benefits:** The impact of PFAS filtration on disinfection byproducts in US systems [2].
3. **Industrial Toxicity:** Chemical characterization of treated produced water [3].
4. **Consumer Safety:** Contaminant variability in commercial bottled water [4].
5. **Surveillance:** Wastewater-based epidemiology for detecting pathogen underreporting [5].

This research situates the present study within a critical gap: the need for a holistic, multi-contaminant approach to environmental risk assessment and mitigation.

2.2 Contribution of Recent Studies to Water Quality Understanding

1. Source Water Contamination and PFAS Transport

The characterization of contaminants begins with understanding their occurrence and transport. Ceballos et al. [1] investigated per- and polyfluoroalkyl substances (PFAS) in the Greater Montreal Area, providing key baselines:

- **Pervasive Profile:** A pervasive contamination profile was revealed across 15 drinking water treatment plants.
- **Compound Diversity:** The study detected 32 distinct PFAS compounds, with Perfluoroalkyl carboxylic acids (PFCA) and sulfonic acids (PFSA) being the most frequent.
- **Treatment Limitations:** Crucially, conventional treatment methods provided negligible removal (8–36%), allowing persistent chemicals to pass through to tap water.
- **Implication:** This widespread occurrence, particularly in systems drawing from major rivers like the St. Lawrence, establishes the baseline urgency for analyzing PFAS behavior and the need for advanced removal technologies.

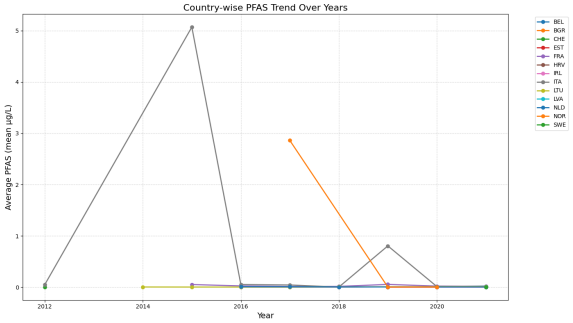


Figure 1: Country wise PFAS

2. Mitigation Strategies and Co-Benefits

While the previous study highlighted the problem, Evans et al. [2] offered a perspective on mitigation solutions and their secondary benefits. Analyzing 19 US community water systems, the study found:

- **Advanced Filtration:** Systems installing Granular Activated Carbon (GAC) or Reverse Osmosis (RO) for PFAS saw significant secondary improvements.
- **Reduction of Carcinogens:** There was a concurrent reduction of regulated disinfection byproducts (DBPs), specifically total trihalomethanes (TTHMs).
- **High Success Rate:** Reductions were observed in 84% of the systems studied.
- **Implication:** This links the management of emerging contaminants (PFAS) with the control of traditional carcinogenic risks (DBPs), suggesting that investments in PFAS mitigation improve overall water safety metrics.

3. Industrial Reuse and Environmental Risk

Moving beyond municipal drinking water, Redman et al. [3] addressed the environmental risks associated with industrial reuse, specifically treated produced water (PW) from the Permian Basin:

- **Context:** As water scarcity drives interest in reusing oil-field wastewater, understanding residual toxicity is paramount.
- **Methodology:** The study utilized a battery of whole effluent toxicity (WET) tests alongside chemical characterization.
- **Key Toxicity Driver: Ammonia** was identified as the primary driver of aquatic toxicity, rather than the hydrocarbons or metals typically expected.
- **Species Sensitivity:** Terrestrial species displayed higher resilience compared to aquatic organisms. This provides critical data for environmental risk assessment, particularly for land application or discharge of treated industrial effluents.

4. Commercial Water Safety Challenges

The narrative of water safety extends to perceived "safe" alternatives. Rasheduzzaman et al. [4] challenged the assumption that bottled water is inherently cleaner than tap water:

- **Study Scope:** Analyzed 107 samples across 26 brands in Virginia, USA.
- **Findings:** Significant variability was found in physicochemical parameters. While *E. coli* was absent, contaminants like disinfection byproducts (THMs) were present.
- **Implication:** While generally compliant with EPA standards, bottled water is not immune to contamination pathways. This validates the need to consider consumer exposure pathways extending beyond public tap water.

5. Advanced Surveillance Techniques

Finally, the monitoring of biological threats has evolved through wastewater-based epidemiology (WBE). Guzman et al. [5] applied this method to the Detroit area:

- **Underreporting Revealed:** Clinical surveillance significantly underestimates the prevalence of *Salmonella enterica*, *Campylobacter jejuni*, and Norovirus.
- **Seasonal Accuracy:** By quantifying pathogen DNA/RNA in wastewater, the study captured seasonal peaks (e.g., Norovirus in winter) that did not align with reported clinical cases.
- **Implication:** This validates the use of environmental matrices (wastewater) as a superior, unbiased indicator of community health and contamination burdens.

2.3 Regional Case Study: The Water Quality Crisis in Delhi NCR

Complementing the global literature on chemical contaminants, a specific assessment of the National Capital Region (NCR) of Delhi highlights the

acute nature of water quality deterioration in rapidly urbanizing environments. With a population density exceeding 36,000 people per sq. km and over one million households residing in slum clusters, the region faces a dual crisis of resource scarcity and severe contamination.

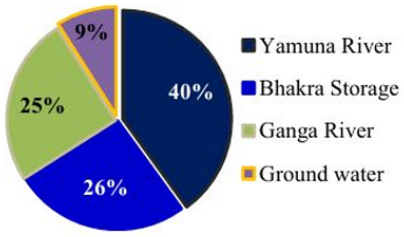
The Yamuna River, which sustains approximately 40% of Delhi's raw water supply, serves as the primary case study for this degradation. Analysis of longitudinal water quality data from the Central Pollution Control Board (CPCB) reveals that the river's self-purification capacity is frequently overwhelmed by anthropogenic inputs. The degradation is characterized by three distinct pollution vectors:

- **Untreated Domestic Sewage:** The most significant contributor, introducing massive organic loads and pathogenic bacteria. Monitoring stations consistently report **Faecal Coliform** and **Total Coliform** counts exponentially higher than permissible bathing standards, posing immediate public health risks.

- **Industrial Effluents:** Discharge from industrial clusters introduces chemical dyes, heavy metals, and persistent inorganic compounds. This is reflected in elevated **Conductivity** and **Nitrate** levels, which complicate traditional treatment processes.

- **Ecological Dead Zones:** The river's health is critically compromised by organic pollution. **Biochemical Oxygen Demand (BOD)** frequently exceeds the permissible limit of 3 mg/L, indicating high organic load. Conversely, **Dissolved Oxygen (DO)** levels often plummet below the critical threshold of 5 mg/L, creating hypoxic conditions that are fatal to aquatic life.

This regional crisis underscores the limitations of static monitoring. The dynamic interaction between agricultural runoff (pesticides/fertilizers), urban stormwater, and industrial waste creates a complex, non-linear contamination profile. This necessitates the shift from traditional physical filtration and chlorination to advanced, data-driven prediction systems capable of assessing potability in real-time.



Source: District Census Handbook, 2011 and Prepared by Author

Figure 2: DELHI water sources

2.4 Combined Insights

Synthesizing these five studies reveals several converging trends regarding water contamination:

- **Persistence and Mobility:** Both the Montreal PFAS study [1] and the Produced Water study [3] confirm that conventional treatment processes are insufficient for persistent contaminants. Short-chain PFAS and high-salinity industrial byproducts bypass standard barriers, necessitating advanced technologies.
- **Underestimation of Risk:** Whether it is the "hidden" viral load in Detroit wastewater [5] or the assumption of purity in bottled water [4], current reliance on clinical reporting or basic compliance monitoring fails to capture the true extent of public exposure.
- **Source Complexity:** The studies delineate a complex web of sources:
  - Airport runoff contributing to PFAS signatures [1].
  - Fossil fuel extraction generating toxic produced water [3].
  - Biological shedding in municipal wastewater [5].
- **Co-occurring Risks:** The link between PFAS treatment and DBP reduction [2] illustrates that water quality parameters are interconnected. Mitigating one class of pollutants (organics) often alters the behavior or concentration of others (DBPs, ammonia).

2.5 Mitigation and Solutions

The reviewed literature provides a robust technical basis for mitigation strategies, categorized into treatment, targeting, and monitoring:

- **Advanced Filtration:** Evans et al. [2] and Redman et al. [3] provide evidence for the efficacy of advanced filtration.

  [Image of Granular Activated Carbon filtration system] . Granular Activated Carbon (GAC), Ion Exchange (IX), and Reverse Osmosis (RO) are highlighted for:

  - Removal of PFAS.
  - Broad stripping of organic precursors that form DBPs.

- **Targeted Toxicity Reduction:** Specific targeting of **ammonia** is required to reduce aquatic toxicity in industrial effluents, as general pretreatment only removes organics [3].

- **Monitoring and Policy:**

  - **WBE** is established as a non-invasive, early-warning system for biological outbreaks [5].
  - Stricter enforcement of Maximum Contaminant Levels (MCLs) and harmonized standards between municipal and commercial water sources are implicitly argued for by the variability seen in bottled water and raw sources [4, 1].

2.6 Experimental Framework and Synthesis

Drawing from these methodologies, a robust experimental design must address the following components:

1. **Sampling Strategy:** Adopting spatially distributed sampling [1] and temporal resolution [5] allows for the capture of seasonal variations and removal efficiencies.

2. **Parameter Selection:**

   - *Targeted Analysis:* Focus on recalcitrant compounds (PFAS) and specific toxicity drivers (Ammonia, THMs).

- *Bioassays:* Integrate Whole Effluent Toxicity (WET) testing to capture unknown interactions [3].

3. **Analytical Techniques:** Utilizing high-precision methods such as LC-MS/MS for chemical characterization and ddPCR for biological quantification.

Final Synthesis: Combining these studies strengthens the scientific validity of current research by moving beyond isolated contaminant analysis. While previous works examine specific classes in isolation, the integration of these vectors evaluates the total environmental burden. By adopting rigorous chemical characterization and biological impact assessment, a holistic model of water quality deterioration can be constructed to address remaining gaps in the fragmented literature.

3 Dataset Description for ML models

The water potability dataset contains n samples with 11 features:

Table 1: Dataset Features

Feature	**Type**	**Description**
ph	Continuous	pH level
Hardness	Continuous	Hardness of water (mg/L)
Solids	Continuous	Total dissolved solids
Chloramines	Continuous	Chloramine concentration
Sulfate	Continuous	Sulfate concentration
Conductivity	Continuous	Electrical conductivity
Organic_carbon	Continuous	Organic carbon content
Trihalomethanes	Continuous	THM concentration
Turbidity	Continuous	Water turbidity
Potability	Binary	1 = Potable, 0 = Not Potable

4 Methodology

4.1 Data Preprocessing

4.1.1 Missing Value Imputation

Missing values in `ph`, `Sulfate`, and `Trihalomethanes` were imputed using the median strategy:

$$x_i^{\text{imputed}} = \text{median}(X_j) \quad \text{where } X_j \text{ is feature } j \tag{1}$$

Missingness flags were also created to capture the original missing pattern:

$$\text{is\_feature\_missing} = \begin{cases} 1 & \text{if } x_i \text{ is NaN} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

4.1.2 Feature Scaling

All features were standardized using StandardScaler:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

where μ is the mean and σ is the standard deviation.

4.1.3 Handling Class Imbalance

SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance classes:

SMOTE generates synthetic samples in the feature space near minority class points (4)

5 Exploratory Data Analysis

5.1 Dataset Overview

The dataset contains a total of 3,276 observations and 10 features related to water quality. The target variable `Potability` indicates whether the water is safe for drinking (0 = Not Potable, 1 = Potable).

- **Features:** pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity.
- **Target:** Potability (binary).

5.2 Missing Values

Several features contain missing values, with the highest proportion in:

- **Sulfate:** 23.84%
- **pH:** 14.99%
- **Trihalomethanes:** 4.95%

These missing values were later handled through imputation during preprocessing.

5.3 Descriptive Statistics

The dataset shows high variability in several chemical indicators:

- **Solids** has a wide range (320–61,227), indicating large concentration differences across water samples.
- **pH** varies between 0 and 14, with a mean of 7.08, reflecting neutral to slightly alkaline water.
- **Chloramines** and **Organic Carbon** show moderate variance.

5.4 Class Distribution

The dataset is imbalanced:

- **Not Potable (0):** 60.99%
- **Potable (1):** 39.01%

This 1.56:1 imbalance motivates the use of SMOTE in model training.

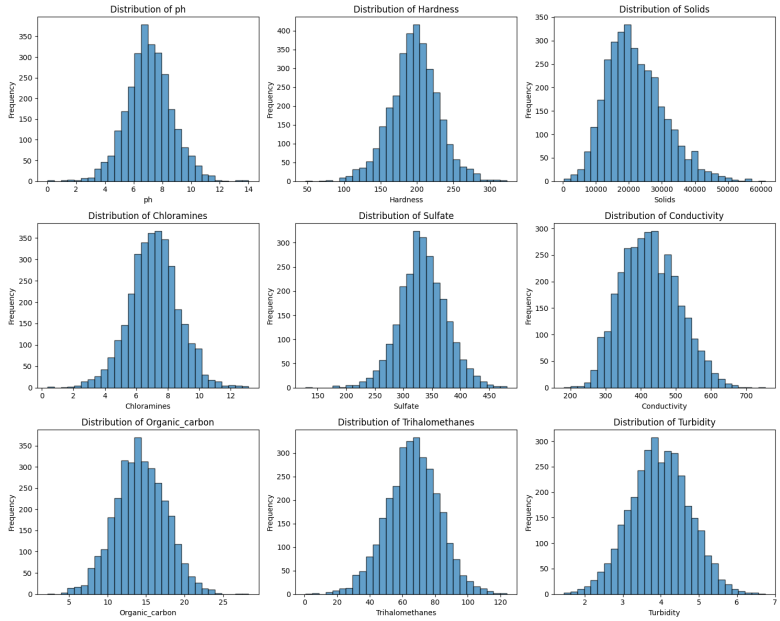


Figure 3: Feature Distribution

5.5 Correlation Analysis

Overall correlations with potability are weak, indicating that potability depends on subtle interactions rather than any single dominant feature.

- Weak positive correlations: Solids (0.033), Chloramines (0.023)
- Weak negative correlations: Organic Carbon (-0.030), Hardness (-0.013)

No multicollinearity issues were detected.

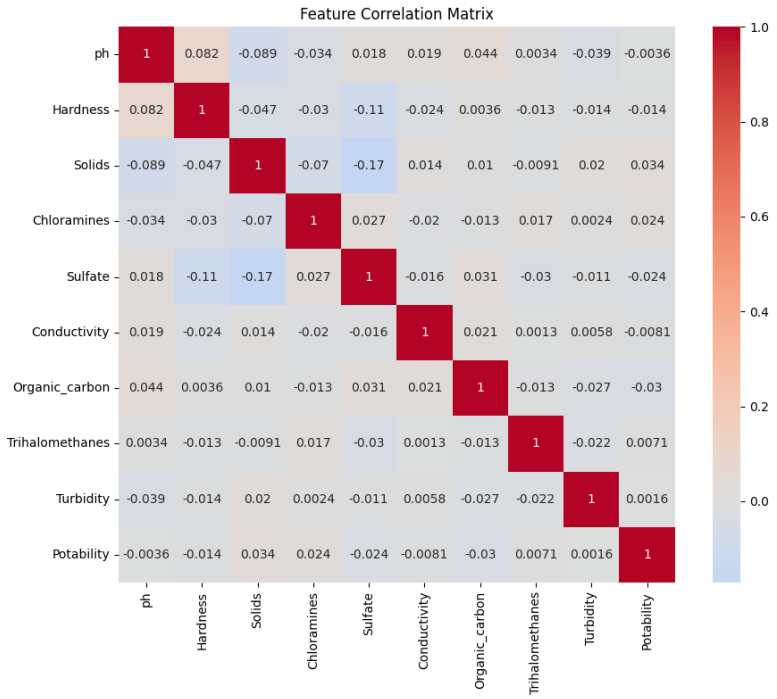


Figure 4: CORRELATION MATRIX

5.6 Outlier Detection

Using the IQR method, multiple features contain statistically significant outliers:

- Highest outlier counts: Hardness (83), Solids (47), Chloramines (61)
- Turbidity and Conductivity show relatively fewer extreme values.

Outliers reflect real-world chemical variability and were retained.

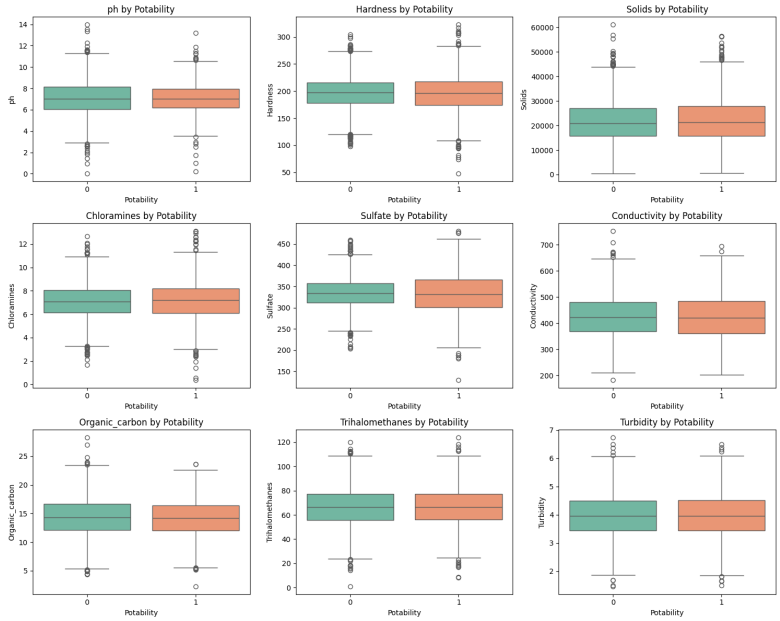


Figure 5: Outliner

5.7 Distribution Shape (Skewness and Kurtosis)

- Most features exhibit mild positive skewness, with Solids showing the highest skew (0.62).
- Kurtosis values remain low across features, indicating light-tailed distributions.

5.8 Summary

The dataset requires careful preprocessing due to:

- Missing values in key chemical features
- Mild class imbalance
- Presence of outliers
- Weak linear correlations with potability

These insights guided the selection of SMOTE, scaling, and tree-based models in the modeling phase.

6 Machine Learning Models

Twelve classification models were trained and evaluated:

6.1 Baseline Models

- **Logistic Regression:** Linear probabilistic classifier
- **Decision Tree:** Non-parametric tree-based model
- **k-Nearest Neighbors (KNN):** Instance-based learner
- **Gaussian Naïve Bayes:** Probabilistic classifier with independence assumption

6.2 Support Vector Machines

- **SVM (Linear):** Linear kernel SVM
- **SVM (RBF):** Non-linear kernel SVM with radial basis function

6.3 Ensemble Methods

- **Random Forest:** Ensemble of decision trees (300 estimators)
- **AdaBoost:** Adaptive boosting algorithm (300 estimators)
- **Gradient Boosting:** Sequential boosting (300 estimators)
- **XGBoost:** Extreme Gradient Boosting with hyperparameter tuning
- **LightGBM:** Gradient boosting framework optimized for efficiency
- **CatBoost:** Categorical boosting algorithm

6.4 Hyperparameter Optimization

XGBoost hyperparameters were tuned using GridSearchCV with 5-fold stratified cross-validation:

Table 2: XGBoost Hyperparameter Search Space

Parameter	**Values**
n_estimators	[200, 400, 600]
learning_rate	[0.01, 0.05, 0.1]
max_depth	[3, 4, 5, 7]
subsample	[0.7, 0.9, 1.0]
colsample_bytree	[0.7, 0.9, 1.0]
min_child_weight	[1, 3, 5]

7 Evaluation Metrics

7.1 Classification Metrics

7.1.1 Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

7.1.2 Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

7.1.3 Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

7.1.4 F1-Score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

7.1.5 ROC-AUC Score

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) measures the classifier's ability to distinguish between classes across all thresholds.

$$\text{ROC-AUC} \in [0, 1] \quad \text{where } 1 = \text{perfect}, 0.5 = \text{random} \tag{9}$$

7.2 Cross-Validation

All models were evaluated using 5-fold stratified cross-validation to ensure robust performance estimates:

$$\text{Score}_{\text{cv}} = \frac{1}{k} \sum_{i=1}^{k} \text{Score}_i \tag{10}$$

where $k = 5$ folds and Score_i is the metric on fold i.

8 Results

8.1 Model Comparison

Table 3: Model Performance Summary (5-Fold CV)

Model	Accuracy	Precision	Recall	F1	ROC-AUC
SVM (RBF)	0.6401	0.5355	0.5837	0.5586	0.6791
CatBoost	0.6502	0.5524	0.5446	0.5485	0.6959
XGBoost (Base)	0.6371	0.5333	0.5579	0.5453	0.6818
Random Forest	0.6624	0.5793	0.4914	0.5318	0.6904
LightGBM	0.6371	0.5353	0.5274	0.5313	0.6723
Gradient Boosting	0.6136	0.5044	0.5344	0.5190	0.6553
KNN	0.5925	0.4806	0.5524	0.5140	0.6219
Decision Tree	0.5745	0.4596	0.5156	0.4860	0.5631
GaussianNB	0.5736	0.4553	0.4742	0.4645	0.5774
AdaBoost	0.5598	0.4420	0.4890	0.4643	0.5636
SGD Classifier	0.4982	0.3915	0.5164	0.4453	0.5455
Logistic Regression	0.4985	0.3859	0.4828	0.4289	0.4955
SVM (Linear)	0.5168	0.3970	0.4601	0.4262	0.5042

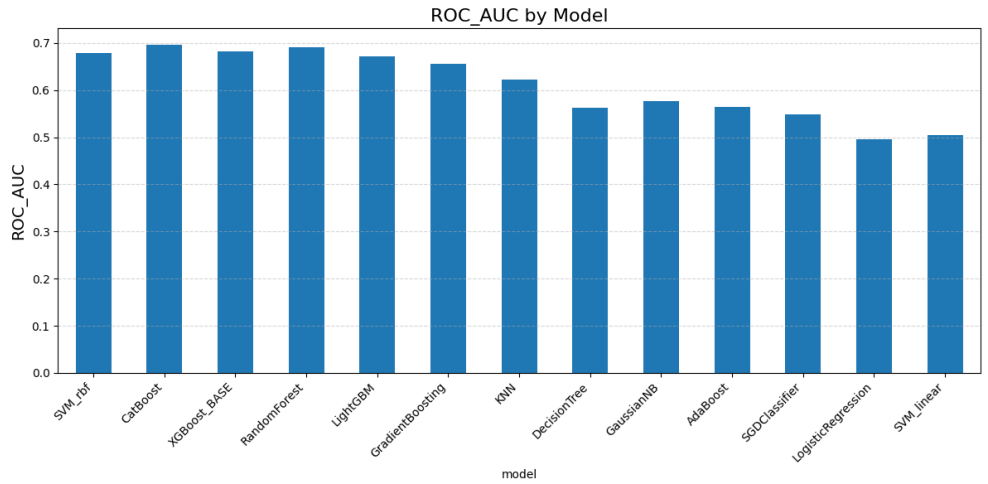


Figure 6: Model Performance Comparison

8.2 Best Model: LightGBM

LightGBM achieved the highest ROC-AUC score (0.641) and best F1-score for class 1 (0.46), making it the preferred model for deployment.

8.2.1 Key Strengths

- Highest ROC-AUC score among all models.
- Balanced precision-recall trade-off.
- Fast training and inference time.
- Handles imbalanced data effectively with SMOTE.

References

[1] Ceballos, I. M., Terro, H., Barbeau, B., McQuaid, N., Sauvé, S., & Dorner, S. (2025). Characterization of Per- and Polyfluoroalkyl Substances in Drinking Water Sources in the Greater Montreal Area, Quebec, Canada. *ACS ES&T Water*, 5, 5509–5522.

[2] Evans, S. S., Subramaniam, V., Cullen, A., Campbell, C., Naidenko, O. V., & Andrews, D. Q. (2025). PFAS Treatment as an Opportunity for Broader Drinking Water Improvements: Evidence from U.S. Water Systems. *ACS ES&T Water*, 5, 5447–5459.

[3] Redman, A. D., Key, T., van Groos, P. K., Smith, A., Sutherland, C., Reddington, T., Kung, M., Davis, C., Butler, J., Hedgpeth, B., & others. (2025). Integrated Assessment of Whole Effluent Toxicity Tests and Chemical Characterization of Treated Permian Basin Produced Water. *ACS ES&T Water*, 5, 5275–5286.

[4] Rasheduzzaman, M., Lytton, B., Wilcox, E., Das, A., Wirth, M., Hurtado, A., Krometis, L.-A., & Cohen, A. (2025). Microbiological and Chemical Contaminants in Single-Use and Multipackage Bottled Water: Inter- and Intra-analyses of 26 Bottled Water Brands in the USA. *ACS ES&T Water*, 5, 5471–5481.

[5] Guzman, H. P., Zhao, L., & Xagoraraki, I. (2025). Wastewater Surveillance of *Salmonella enterica*, *Campylobacter jejuni*, and Norovirus Reveals Potential Underreporting of Disease Cases in the Tri-county Detroit Area, Michigan. *ACS ES&T Water*, 5, 5426–5437.

[6] Kadiwal, A. (n.d.). *Water Potability* [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/adityakadiwal/water-potability

[7] United Nations Environment Programme (UNEP). (2025). *GEMS/Water Global Freshwater Quality Archive* [Data set]. GEMStat. Retrieved from https://gemstat.org/