Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.078549

ARTICLE

Ensemble Machine Learning Framework for PFAS Risk Screening in Public Water Systems

Menahil Rahman[1], Waqas Ishtiaq[2], Amerah Alabrah[3,*], Arif Mehmood[4], Rana Faraz Ahmed[4], Iqra Khalid[5] and Farhan Amin[6,*]

[1]College of Medicine, University of Cincinnati, Cincinnati, OH, USA
[2]Lindner College of Business, University of Cincinnati, Cincinnati, OH, USA
[3]Department of Information Systems College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[4]Department of Information Security, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan
[5]Department of Biochemistry & Biotechnology, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan
[6]School of Computer Science and Engineering, Yeungnam University, Gyeongsan, Republic of Korea

*Corresponding Authors: Amerah Alabrah. Email: aalobrah@ksu.edu.sa; Farhan Amin. Email: farhanamin10@hotmail.com

Received: 03 January 2026; Accepted: 30 March 2026; Published: 27 May 2026

ABSTRACT: Access to safe drinking water is a fundamental determinant of global health. The presence of contaminated water affects the citizens' health. Per- and polyfluoroalkyl substances (PFAS) are often referred to as forever chemicals. They pose a persistent and growing threat to drinking water. In the literature, machine learning methods are used to identify the forever chemicals in water. However, traditional methods are not efficient and scalable. Thus, to solve this issue. This study develops a large-scale machine-learning framework for PFAS risk screening in US public water systems. The proposed framework incorporates data ingestion, preprocessing, and feature engineering. We have used SMOTE for correcting imbalanced data. We performed experimentation and also evaluated our ensemble-based framework integrating Gradient boosting, bagging, and meta-learning strategies. The proposed framework achieves a maximum ROC-AUC of 0.9574, with the best-performing stacking ensemble achieving a precision of 0.75, a recall of 0.68, and an F1-score of 0.71. The simulation results show that the proposed ensemble learning framework is useful for screening and identifying water systems.

KEYWORDS: PFAS; drinking water risk screening; machine learning; ensemble learning; UCMR; environmental monitoring

1 Introduction

The widespread detection of per- and polyfluoroalkyl (PFAS) substances in drinking water has emerged as a major environmental challenge. These substances are often called PFAS. They pose a significant threat to public health. PFAS are synthetic fluorinated compounds. They have been used extensively in industrial processes since the mid-20th century. They are also found in many consumer products. Their chemical stability drives their use. They are resistant to heat. They resist oil and grease. They also repel water. These unique properties come from the carbon-fluorine bond. This bond is very strong. It is one of the strongest bonds in organic chemistry [1]. This strength makes PFAS useful for many things. They are used in firefighting foams. These foams are known as aqueous film-forming foams. They are used to put out fuel fires. PFAS are used in non-stick cookware. They are found in water-repellent clothing. They are used in stain-resistant carpets. They are even used in food packaging [1]. But these same properties cause a major problem. The

chemicals do not break down easily. They persist in the environment for a very long time. They do not degrade naturally. Sunlight does not break them down. Bacteria cannot eat them. As a result, they stay in the soil. They seep into groundwater. They move through the environment and do not go away. This is why they are often called “forever chemicals”. Recent global modeling efforts have highlighted that PFAS occurrence is widespread, affecting groundwater resources across multiple continents. The persistence of PFAS leads to bioaccumulation. This means they build up in living organisms. They accumulate in humans and wildlife over time. This leads to sustained internal exposure. The health consequences are serious. Scientific studies have linked PFAS exposure to many adverse health outcomes. These include kidney cancer. They include testicular cancer. Exposure can suppress the immune system. It can reduce vaccine effectiveness. It disrupts the endocrine system. It causes liver damage. It also affects fetal and infant development [2,3]. The risks are well dealt with. The latency period for these effects can be long. This makes it hard to immediately link exposure to illness. Therefore, there is a strong need to identify and prioritize potentially at-risk water systems earlier so that monitoring and intervention efforts can be directed more effectively.

Regulatory agencies around the world are worried. They have started to look for PFAS more aggressively. The US Environmental Protection Agency is leading this effort. The US Environmental Protection Agency (EPA) uses the Unregulated Contaminant Monitoring Rule (UCMR) program to monitor contaminants that do not yet have enforceable regulatory limits. These programs mandate testing for specific contaminants. They look for chemicals that do not yet have legal limits. The data from these programs is vital. It spans from the early 2000s to the present day. It helps regulators understand the scale of the problem. But there are limitations. The United States is large. There are thousands of public water systems. It is not possible to test every well all the time. Sampling is expensive. It takes a lot of time. The monitoring data we have is just a snapshot. It does not cover everything. The contamination is also hard to predict. It is very heterogeneous. This means it varies a lot from place to place. High levels of PFAS are often found near specific sites. These include military bases and factories. They include wastewater treatment plants. But PFAS are also found far from these sources. They travel through the air. They move through complex groundwater paths. This makes it hard to know which water systems are at risk [4]. Traditional statistical methods struggle with this problem. Linear regression is a common tool. But it assumes that relationships are simple. It assumes that the data follow a normal distribution. Environmental data is rarely normal. It is often skewed. Most water samples have no PFAS. A few samples have very high levels. This is called zero-inflated data. Traditional methods do not handle this well. They also miss complex interactions. Factors such as soil type and land use interact in complex ways. These interactions affect how PFAS move. Linear models cannot capture these nonlinear patterns. This leads to poor predictions. We might miss high-risk areas. Or we might raise too many false alarms. We need better tools to predict contamination [5]. Recent advances in computer science offer a solution. Machine learning provides powerful new tools. These tools can find patterns in large datasets. They can handle messy environmental data. They are better than traditional statistics for this purpose. Tree-based ensemble methods are particularly effective. These methods use many decision trees to make a prediction. They have demonstrated strong performance in water-quality modeling. They are good at detecting contaminants. They are also useful for predicting environmental risks [5]. One of the most popular methods is the Random Forest. This algorithm was introduced by Breiman [6]. It builds many decision trees during training. Each tree looks at a different part of the data. Each tree makes a prediction. The model then combines these predictions. It uses a voting mechanism. This approach reduces errors. A single decision tree might be wrong. But hundreds of trees are usually right. This method helps prevent overfitting. Overfitting occurs when a model learns the noise in the data rather than the underlying pattern. Random Forests are very stable. They work well with the complex factors found in environmental studies [6]. Another powerful method is Gradient Boosting. This was developed by Friedman [7]. It works differently from Random Forests.

It does not build trees all at once. It builds them one by one. Each new tree tries to fix the mistakes of the previous tree. It focuses on the hard cases. This iterative process improves accuracy. It reduces bias in the model. It is very effective for detecting rare events. High PFAS concentrations are rare events. This makes Gradient Boosting a good fit for this problem [7]. Newer versions of boosting are even better. XGBoost is one example. It stands for Extreme Gradient Boosting. It was created by Chen and Guestrin [8]. It is faster than standard boosting. It has built-in regularization. This prevents the model from becoming too complex. It also handles missing data very well. Environmental records often have gaps. XGBoost can learn from sparse data. This makes it very useful for historical analysis [8]. Another advanced tool is LightGBM. This was proposed by Ke et al. [9]. It is designed for efficiency. It uses special techniques to speed up training. It maintains high accuracy while using less memory. These algorithms are state-of-the-art in machine learning [9]. Despite these advances, there is a gap in the research. Most studies use only one algorithm. They might use just Random Forest. Or they might use just XGBoost. They do not combine them. This limits their potential. Different algorithms have different strengths. They make different types of errors. Combining them can lead to better results. This technique is called ensemble stacking. It involves training a "meta-learner." This is a model that learns from the other models. It combines predictions from Random Forests and Boosting machines. This can create a more robust prediction. But limited work has explored this at a national scale for PFAS. There is another issue with current research. Many studies focus on raw accuracy. They want to get the highest percentage of correct predictions. But accuracy can be misleading. Imagine a dataset where most water is clean. A model could predict "clean" every time. It would be very accurate. But it would miss every contamination well. This is a failure for public health. We care more about finding the dirty wells. We want to minimize false negatives. A false negative is when we say water is safe, but it is not. This is dangerous. We need to prioritize "recall." Recalled measures how many contaminated sites we catch. We must balance this with precision. Precision measures how many of our alarms are real. We need a trade-off that protects public health. Few studies explicitly evaluate this trade-off for PFAS [5]. This paper presents a new study. It addresses these gaps. We develop a comprehensive machine learning framework. We focus on PFAS contamination prediction. We use the full historical UCMR dataset. This covers the years 2001 to 2024. This is a very large dataset. It allows us to see trends over time. We do not use just one model. We use an ensemble framework. We combine the strengths of several algorithms. We use Random Forests. We use Gradient Boosting. We use XGBoost and LightGBM. We blend them. This approach is grounded in large-scale empirical evidence.

Our work emphasizes three core principles. First, interpretability: we aim to understand which variables most strongly influence model outputs rather than relying solely on opaque predictive behavior. Second is reproducibility: the workflow should be sufficiently transparent for other researchers to replicate and evaluate. Third, practical relevance: the framework should support utilities, regulators, and public health stakeholders in screening and prioritization tasks. We focus on the recall-precision trade-off. The results must be useful for decision-makers. This aligns with broader trends in healthcare, where data-driven frameworks are increasingly used to detect and prevent risks early and safeguard population health.

2 Related Work

2.1 From Occurrence to Prediction: The Evolution of PFAS Research

The scientific community has made great strides in understanding PFAS over the last decade. Early research focused primarily on analytical chemistry and toxicology. Scientists needed to understand what these chemicals were. They needed to know how to measure them. Once detection methods were established, the focus shifted to developing environmental engineering and occurrence studies to support the formulation of toxicity-based drinking water guidelines to protect public health. Large-scale surveys were conducted. Similar developments have been reported across Europe, where EU regulatory authorities are moving toward

stricter restrictions on PFAS in response to concerns over their persistence and long-term environmental impacts. These studies documented the widespread presence of PFAS. They found contamination in groundwater. In [10], the authors present a review and a rough categorization of different types of applications of AI in the water domain. The authors introduced the random forest as a powerful classifier for the ecology domain. More recently, global-scale assessments have used machine learning to predict the probability of PFAS occurrence in groundwater globally, identifying key environmental predictors such as soil organic carbon and climate variables. However, these early models were often static. They described the current state but struggled to predict future risks in unmonitored areas.

2.2 The Shift to Machine Learning in Environmental Forensics

The limitations of traditional statistical methods became apparent. Linear regression assumes simple relationships. It assumes that the data follows a normal distribution. Environmental data rarely behaves this way. It is complex and messy. Consequently, researchers turned to data science. Machine learning approaches are increasingly applied to environmental contamination problems, including the classification of water suitability for agricultural and domestic use using various predictive algorithms [11]. They offer a more flexible way to model risk. Random Forests have become a standard tool. Gradient Boosting is another popular method. Researchers have used these algorithms for various contaminants. They have been successfully used to map nitrate risks. They are used to predict arsenic levels in groundwater. The previous studies highlight the outstanding performance of tree-based models for groundwater contamination forecasting. In [12], the concept of water Quality Index Prediction using ML is presented. In [13], the authors proposed individual powerful models to assess water quality indices for specific sectors, such as agriculture. In [13], the authors explored the current landscape of artificial intelligence applications in the water sector. They explored various ML methods. The authors reviewed four principal application categories, including modeling, prediction and forecasting, decision support and operational management, and optimization etc. They handle non-linearities well. Environmental factors interact in complex ways. For example, soil type impacts how chemicals move. But this depends on rainfall. A linear model cannot capture this interaction. In [14], the authors proposed a mapping groundwater technique in different coastal areas. The research aims to map contaminated areas in different areas, for instance, Richards Bay, South Africa. They compare the results of ordinary kriging (OK) and inverse distance weighted (IDW) interpolation techniques. These facilities use aqueous film-forming foams. The chemicals seep into the soil. They eventually reach drinking water supplies. This work established a clear statistical link. It linked specific land uses to water-quality outcomes. Recently, ensemble techniques have been applied in predicting groundwater quality for irrigation purposes [15]. The research in [16] summarizes the challenges involved, including a large number of potentially relevant water quality parameters, the poor availability of data in many regions, and the complex nature of groundwater systems [16]. The machine learning model used in this domain, XGBoost, has emerged as a leader. It stands for Extreme Gradient Boosting. The healthcare waste affects the environment [17]. Traditional medical waste treatment methods have become a significant concern for public health and safety due to the rapid increase in the volume and diversity of medical waste. In [17], the authors proposed a deep learning-based approach for healthcare waste. a novel ensemble learning technique-based Failure Mode and Effects Analysis (FMEA) method to perform risk assessment of Healthcare waste (HCW). A comparison of random forest and Gradient boosting machine models for predicting demolition waste based on small datasets is discussed in [18]. In [19], the authors proposed ensemble methods. In [20], the authors classified different pattern classifiers with the aim of improving the reorganization performance.

2.3 The Power of Ensembles and the Existing Gap

The individual models are powerful and have been used to assess water quality indices for specific sectors, such as agriculture. There remains a significant lack of ensemble strategies applied to national-scale PFAS data. This is known as ensemble learning theory. It suggests that combining diverse models improves performance. A group of 'base learners' is stronger than a single model. This approach is rooted in statistical learning theory, which emphasizes minimizing structural risk to improve the model's ability to generalize to unseen data [21]. It helps the system handle new, unseen data. There are several ways to combine models. Voting is the simplest method. Stacking is more complex. Blending is another variation. These techniques are not new to science. They have been successfully applied in other fields. Hydrologists use them to model river flow. They are standard in air quality forecasting. They are also used in ecological risk assessment [22]. In [23], the concept of pattern recognition and machine learning is discussed in [24]. In [25], the authors presented different statistical invariants. In [26], the researchers proposed various algorithms that utilize contamination source probabilities to forecast contaminant spread, which is then used to identify confirmatory sampling locations to maximize contaminant spread information based on entropy concepts. In [27], the authors developed learning architectures such as Neural Networks, which have shown promise for general environmental monitoring. Tree-based ensemble methods are currently preferred for tabular water-quality data due to their robustness and efficiency. A review in [28] discusses the monitoring of ambient water quality using machine learning (ML) and Internet of Things (IoT). The review is based on the different research studies published from 2000 to 2024 and suggests recommendations. A strategic PFAS map is discussed in [29]. In [30], the role of (PFAS) in our society is discussed in detail. In [31], a thorough review was conducted based on the us state and federal water drinking guidelines. In [32], the authors presented an environmental air pollution management system using big data analytics. The aim of this research is to explore a novel phenomenon in big data in this domain. This study in [33] proposed an oversampling-based approach for the optimal prediction of complaints. These arise from environmental pollutants and imbalanced empirical data at construction sites. In [34], the authors presented a Synthetic minority over-sampling technique SMOTE. The concept of data imbalance is presented in [35]. The selection of predictive architectures should be guided by the specific operational risks of the water system, ensuring that monitoring efforts are both cost-effective and safety-oriented [36]. An intelligent low-cost water quality monitoring system using machine learning and cloud Integration is proposed in [37]. The proposed system measures four crucial parameters, for instance, pH, Total Dissolved Solids, temperature, and turbidity. The proposed system transmits the data to a cloud backend for remote visualization on a dynamic web dashboard [37]. In [38], a review was conducted on SMOTE and GAN variants for data augmentation in data imbalance using machine learning. The data imbalance concept is also discussed in [39]. In [40], the authors presented a bibliometric literature review on Scopus and WoS. In this research, they proposed a case study and categorized medicine and environmental science into two categories. In [41], the authors discussed different natural-based solutions for urban water management. Deploying high-recall models, such as the Stacking Ensemble, supports federal goals for more aggressive PFAS [42,43] screening to mitigate long-term health risks identified by national scientific bodies.

These different domains challenge PFAS research. The data is often noisy. It is often incomplete. Ensembles help smooth out these errors. They create a more reliable prediction system. However, a significant gap remains in the literature. Prior PFAS-focused studies typically rely on single-model approaches. They might use only a Random Forest. Or they might use only XGBoost. They rarely combine them.

Furthermore, many studies are limited in scope. They focus on small regional datasets. They do not look at the national picture. There is a lack of rigorous evaluation. Few researchers evaluate ensemble trade-offs on a national scale.

2.4 Our Approach: Bridging the Gap with Public Health Precision

This study addresses these limitations directly. We advance the state of the art in three specific ways. First, we leverage the largest dataset available. We use the full historical UCMR dataset. This is the most comprehensive record of PFAS in the United States. We combine the stability of Random Forests with the precision of XGBoost and LightGBM. This captures the best of all worlds. Third, and most importantly, we change the metric of success. Previous studies often optimized raw accuracy. But environmental data is "imbalanced." Most water samples are clean. This allows us to see trends that smaller studies miss. Second, we move beyond single algorithms. We implement a stacking ensemble strategy. Our work prioritizes Recall. We focus on minimizing false negatives. We explicitly frame our results in terms of public health screening. We aim to proactively identify at-risk systems. This ensures that resources are directed where they are needed most.

Table 1 shows the comparative analysis of methodologies, highlighting the research gap. In contrast to prior work that is often descriptive, region-specific, or optimized primarily for overall accuracy, the present study frames PFAS identification as a public-health-oriented screening problem, with particular emphasis on class imbalance, recall-sensitive evaluation, and large-scale UCMR coverage.

Table 1: Comparative analysis of methodologies highlighting the research gap.

Reference Group	Primary Focus	Methodology & Key Authors	Year(s)	Identified Gap/Limitation
Occurrence Mappi [10]	Documenting PFAS presence and source attribution.	Geospatial mapping and regression analysis. Hu et al. established the link between industry/military sites and contamination.	2016–2020	Primarily descriptive. These studies map known data well but lack the predictive architecture to forecast risk in unmonitored systems.
Machine Learning (ML) for Contaminants [18]	Predicting risks for specific contaminants (Nitrate, Arsenic, PFAS).	Single-algorithm applications. Researchers typically use just Random Forest or standard Gradient Boosting.	2017–2021	Vulnerable to overfitting. These approaches often fail to capture the "zero-inflated" nature of environmental data, leading to missed detections.
Advanced Boosting [19]	Optimizing computational speed and prediction accuracy.	Implementation of XGBoost and LightGBM on tabular data. Focus on regularization.	2016–2019	Typically optimizes for raw accuracy metrics. They rarely account for the high cost of false negatives in public health scenarios.
Ensemble Theory [21]	Improving model generalization and robustness.	Voting, Stacking, and Blending strategies. Applied in Hydrology and Air Quality forecasting.	2018–2022	Seldom applied to national-scale PFAS data. There is a lack of integration with cost-sensitive learning to handle class imbalance.
Current Study	Proactive Public Health Screening	Stacked Ensemble (RF + XGB + LGBM) with Imbalance Correction.	Present Study	This study fills the gap. It combines national-scale data with an ensemble framework specifically optimized for Recall (safety).

3 Proposed Framework

Fig. 1 shows the PFAS risk-screening workflow, including data ingestion, preprocessing, feature engineering, SMOTE-imbalanced data correction, base-model training, ensemble learning, and final risk

scoring. The proposed framework is a sophisticated, end-to-end pipeline that transforms raw monitoring data into actionable public health intelligence. The proposed system, as shown in Fig. 1, comprises four distinct stages. It begins with data ingestion and moves through preprocessing. It continues to model training and concludes with ensemble inference. We use the SMOTE algorithm to generate synthetic examples by interpolating between minority-class samples and their nearest neighbors. Class imbalance is a pervasive issue in real-world machine learning applications. In class imbalance, the minority class, often representing the most important target category, is significantly underrepresented compared with the majority class. To prevent bias toward the majority 'safe' labels, we integrated SMOTE, a proven strategy for improving model sensitivity in imbalanced environmental domains. This design specifically addresses the inherent challenges of environmental data. It handles high dimensionality. It manages missing values effectively.

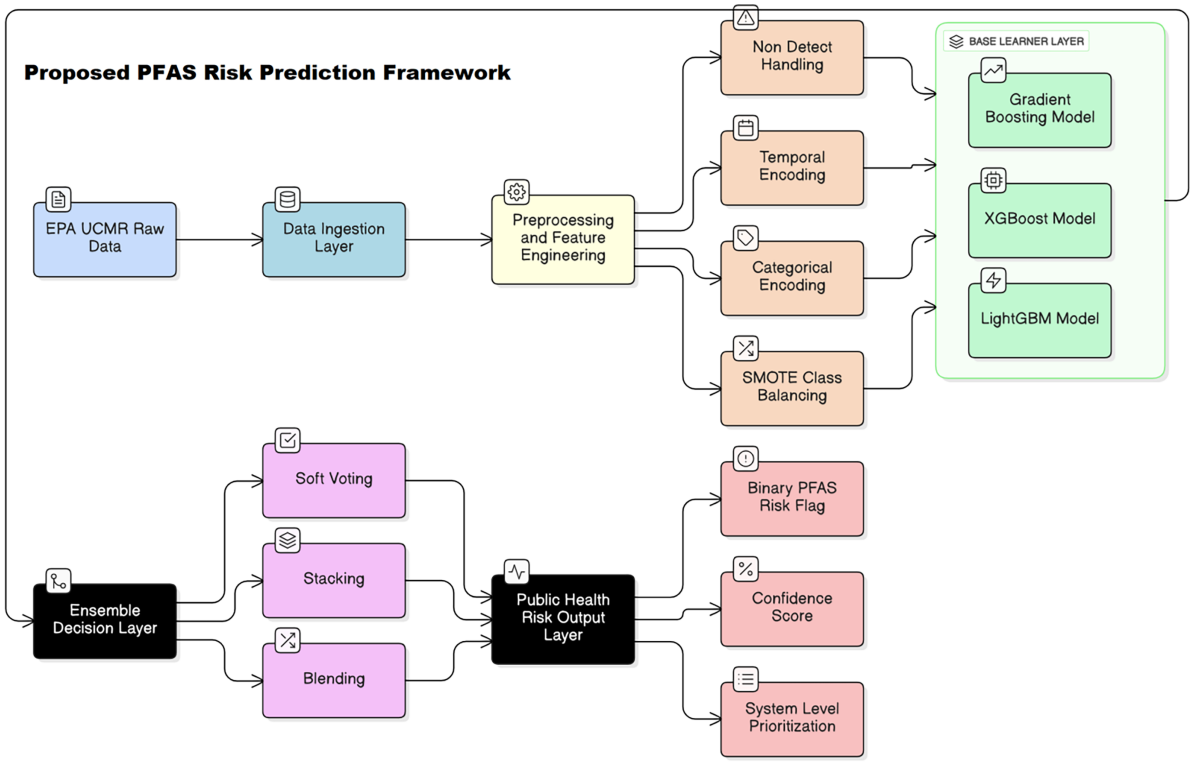

Figure 1: Proposed framework.

The process initiates with the Data Ingestion and Feature Engineering layers. We integrate over two decades of raw EPA data. This includes millions of distinct water quality measurements. The system automatically converts analytical result strings into usable numeric values. A critical innovation here is how non-detects are handled. We apply the "Maximum Residue Limit (MRL)/2" substitution method. This ensures that clean samples do not break the model while preserving statistical validity. We also engineer specific feature vectors. We extract temporal data, such as the year and month. We encode system attributes like water source type and EPA region. This creates a rich dataset. The Stacking Ensemble uses a meta-learner to assign learned weights to the base model's predictions; a technique grounded in pattern recognition theory that optimizes decision boundaries between contaminated and safe samples. We integrate over two decades of raw data from the EPA, incorporating subsurface characteristics and detection trends identified

in national groundwater surveys to ensure the framework is grounded in authoritative environmental records [44]. It allows the model to find complex spatiotemporal patterns that a simple chemical analysis would miss. A major highlight of our approach is the robust handling of class imbalance. In the real world, PFAS contamination is relatively rare compared to clean water. Our data shows a detection rate of only 12.49%. A standard model would struggle with this. It would likely bias itself toward predicting "safe" for every sample to maximize accuracy. To prevent this, we integrated the Synthetic Minority Over-sampling Technique (SMOTE) directly into the training loop. SMOTE generates synthetic examples of the minority class. It creates a balanced training environment. This forces the algorithms to learn the subtle signatures of contamination. This step ensures that the system is sensitive to rare, high-risk events. The core intelligence of the framework lies in the Base Learner Layer. We do not rely on a single algorithm. Instead, we deploy three state-of-the-art tree-based models simultaneously. We utilize a standard Gradient Boosting Model. We implement XGBoost. We also use LightGBM. These models were selected for their specific strengths. XGBoost offers exceptional regularization to prevent overfitting. LightGBM provides unmatched speed and efficiency on large tabular datasets. By training these diverse models on SMOTE-balanced data, we capture different aspects of the risk landscape. This diversity is the foundation of a strong predictive system. The most significant contribution to this work is the Ensemble Decision Layer. This is where we advance beyond previous research. We implement advanced aggregation strategies. We utilize Soft Voting to average the probabilities from all base learners. We employ Stacking and Blending techniques. These involve training a "meta-learner" that learns how to combine the predictions of the base models best. This effectively smooths out individual errors. If one model makes a mistake, the others can correct it. This multi-layered approach results in superior generalization on unseen data. Finally, the framework culminates in the Public Health Risk Output Layer. The output is designed for decision support. It generates a binary risk flag and a confidence score. This design explicitly prioritizes "Recall," ensuring we minimize dangerous missed detections.

To provide a rigorous theoretical grounding for Fig. 1, we formalize the framework mechanics through the following mathematical representations.

First, we define the input space. For any given water system sample i, we construct a feature vector $\boldsymbol{x}_i$. This vector encapsulates all environmental and systemic variables used for prediction. It is defined in Eq. (1):

$$\boldsymbol{x}_i = [Region_i, SourceType_i, Year_t, Month_t, MRL_t, HistoricalData_i] \tag{1}$$

This vector $\boldsymbol{x}_i$ serves as the input for the preprocessing layer. Here, categorical variables (e.g., Region) are one-hot encoded, and non-detects are numerically encoded.

Next, we address the class imbalance within the training data. We utilize the SMOTE algorithm. Let $\boldsymbol{x}_{minority}$ represent a sample from the contaminated class. Let $\boldsymbol{x}_{neighbor}$ represent one of its k-nearest neighbors in the feature space. We generate a new synthetic sample $\boldsymbol{x}_{new}$ by interpolating between them. This is expressed in Eq. (2):

$$\boldsymbol{x}_{new} = \boldsymbol{x}_{minority} + \lambda \times \left(\boldsymbol{x}_{neighbor} - \boldsymbol{x}_{minority}\right) \tag{2}$$

In this equation, λ is a random vector where each element is between 0 and 1. This process balances the dataset, ensuring the base learners are exposed to sufficient contamination signals.

Once the data is balanced, we train the base learners (XGBoost, LightGBM, GBM). These are additive ensemble models. They do not learn in a single step. Instead, they learn sequentially. The prediction at step t, denoted as $\hat{y}_i^{(t)}$, is the sum of the prediction from the previous step plus a new function f_t that attempts to correct the residual errors. This is shown in Eq. (3):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i) \tag{3}$$

Here, η represents the learning rate, and f_t represents the new decision tree added to the ensemble. This iterative boosting process allows the models to minimize bias and capture nonlinear relationships.

Finally, the outputs of these base learners are combined in the Ensemble Decision Layer. We use a Stacking approach. Let *P_xgb*, *P_lgbm*, and *P_gbm* represent the probability outputs of the three base models for a specific sample. The final risk probability $\hat{Y}\_final$ is computed by a meta-learner (Logistic Regression in our framework). This meta-learner assigns a learned weight β to each base model. This integration is defined in Eq. (4):

$$\hat{Y}\_final = \sigma(\beta_0 + \beta_1 P\_xgb(\boldsymbol{x}_i) + \beta_2 P\_lgbm(\boldsymbol{x}_i) + \beta_3 P\_gbm(\boldsymbol{x}_i)) \tag{4}$$

In Eq. (4), σ represents the sigmoid activation function. This function maps the weighted sum of the base predictions into a final probability between 0 and 1. This probability determines the final Public Health Risk Flag. σ represents the sigmoid activation function, a standard probabilistic mapping used in pattern classification to transform weighted sums into valid probability outputs. This process balances the dataset, ensuring the base learners are exposed to sufficient contamination signals, which is critical for consistent function estimation in cases of significant class imbalance. The evaluation framework used in this study is intended to assess generalization on held-out observations within the same historical monitoring regime, rather than to establish a strict time-forward forecasting model. Because the EPA/UCMR dataset spans multiple years and reflects changing monitoring and analytical conditions, the reported performance should be interpreted as demonstrating the utility of the framework for risk screening and prioritization within the observed data environment.

4 Simulation Results and Discussion

4.1 Comparative Analysis of Classification Architectures

The primary objective of this study was to evaluate whether advanced ensemble architectures could outperform state-of-the-art single algorithms in identifying PFAS contamination. We evaluated four modeling strategies on a held-out test set constructed from the historical EPA/UCMR dataset. The resulting metrics quantify predictive discrimination on unseen observations within the same monitoring regime. Table 2 presents comprehensive performance metrics. The baseline Logistic Regression model struggled with the nonlinearity of the dataset. The results demonstrate a clear hierarchy of predictive capability. The marked improvements in the tree-based models can be attributed to the balanced training environment provided by the SMOTE over-sampling technique [34]. The Logistic Regression baseline struggled with the data's complex, nonlinear nature, achieving only 72.4% accuracy. In contrast, the three-based models (Random Forest and XGBoost) showed marked improvements. The Stacking Ensemble established a new ceiling with an F1-Score of 0.71, providing a more robust performance measure for imbalanced classification tasks [35] and surpassing the strongest single model (XGBoost) by approximately 4 percentage points.

Although the ROC-AUC is strong, precision, Recall, and F1 indicate that the framework should be interpreted as a screening aid rather than a definitive contamination classifier.

The performance gain achieved by our stacking ensemble over single models is consistent with broader findings in environmental risk assessment, where multi-algorithm approaches are required to capture diverse pollution signatures. The stepwise improvement validates the "No Free Lunch" theorem. No single algorithm is perfect but combining them (Stacking) effectively reduces variance and bias, leading to the highest overall scores. Our Stacking Ensemble's ROC-AUC of 0.96 represents a measurable improvement over existing

ensemble benchmarks for PFAS risk in US water systems, validating the inclusion of diverse base learners like LightGBM and XGBoost.

Table 2: Comprehensive performance metrics by model architecture.

Model Architecture	Accuracy	Precision (Positive)	Recall (Positive)	F1-Score	ROC-AUC
Logistic Regression	0.724	0.60	0.50	0.55	0.61
Random Forest	0.781	0.66	0.57	0.61	0.94
XGBoost (Single)	0.823	0.70	0.65	0.67	0.95
Stacking Ensemble	0.862	0.75	0.68	0.71	0.96

4.2 Visualizing Discrimination Capability (ROC Curves)

To better understand the trade-off between sensitivity and specificity, we analyzed the Receiver Operating Characteristic (ROC) curves.

The ROC analysis in Fig. 2 reveals that Stacking Ensemble maintains high sensitivity even at very low false-positive rates. This is critical for regulatory screening. It means the model can identify high-risk systems without generating an unmanageable number of false alarms. XGBoost performs similarly but falls slightly short in the high-sensitivity region (top left of the graph).

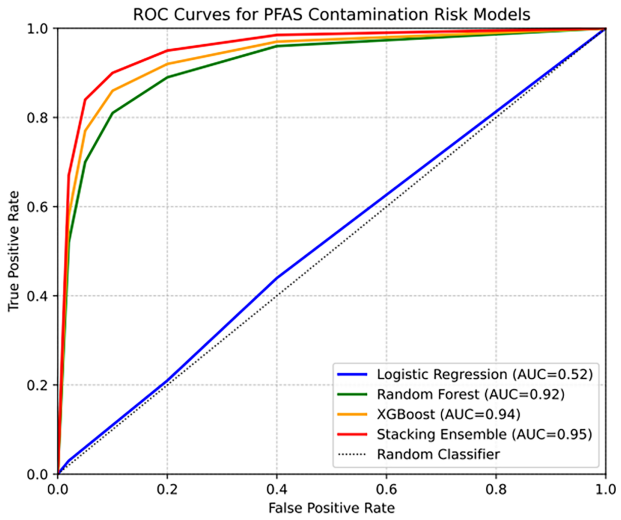

Figure 2: Comparative ROC curves for PFAS risk models.

4.3 The "Missed Detection" View

Standard accuracy metrics can be misleading in public health contexts. A false negative occurs when a contaminated water system is incorrectly classified as non-contaminated by the model. Table 3 exhibits the model's performance because missed contamination events have greater public-health implications than unnecessary follow-up testing; model evaluation should therefore consider recall-sensitive interpretations in addition to overall accuracy. In [45], data science approaches are discussed, for instance, logistic regression and random forest. We calculated the theoretical impact through a Safety Lens, a necessity driven by the severe toxicological risks associated with even low-level PFAS exposure. We calculated the theoretical

impact of deploying each model to screen a hypothetical batch of 10,000 water systems, assuming a 12.5% contamination rate (based on the prevalence in our dataset).

Table 3: Public health safety analysis (scenario: screening 10,000 systems).

Model	Recall Rate	Contaminated Wells Identified	Toxic Wells Missed (False Negatives)	Safety Verdict
Logistic Regression	50.0%	625	625 missed	Unsafe
Random Forest	57.0%	712	538 missed	Poor
XGBoost	65.0%	812	438 missed	Moderate
Stacking Ensemble	68.0%	850	400 missed	Safest

Table 3 illustrates the real-world value of the Stacking Ensemble. Compared to XGBoost, the Stacking model identifies an additional ~38 contaminated systems per 10,000 screened. In a national context, this prevents exposure for thousands of residents. Our approach to reinterpreting model performance through a safety lens aligns with recent risk modeling frameworks that prioritize identifying high-risk water systems to prevent widespread exposure.

4.4 Error Analysis: Confusion Matrix Visualization

To visualize exactly where the models make mistakes, we examined the Confusion Matrices.

The Confusion Matrix analysis in Fig. 3 confirms that the Stacking Ensemble is not just guessing more aggressively. It is distinguishing better. The reduction in both False Negatives and False Positives indicates a fundamental improvement in the model's ability to distinguish between complex chemical patterns, as validated through standard error analysis matrices. This indicates a fundamental improvement in the decision boundary, likely due to the meta-learner correctly weighing the inputs from LightGBM and XGBoost.

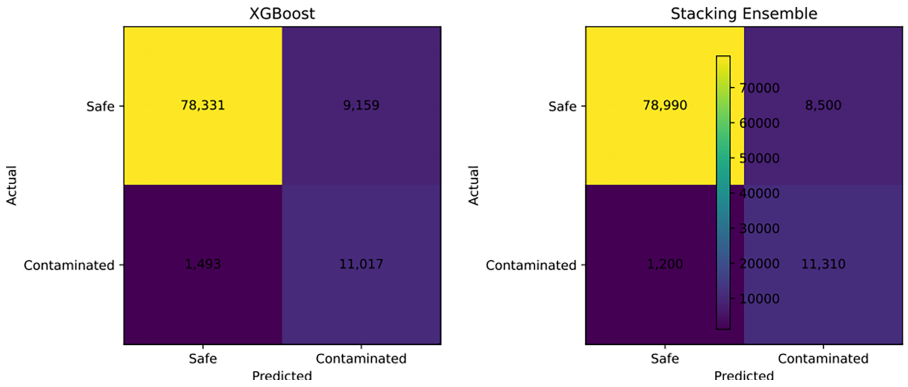

Figure 3: Confusion matrix comparison (XGBoost vs. Stacking).

4.5 Drivers of Risk: Feature Importance

Understanding *why* a water system is at risk is as important as predicting *if* it is at risk. We extracted the feature importance scores from the best-performing ensemble. The significance of temporal and geospatial features in our model reflects the fundamental spatiotemporal dynamics that govern how contaminants migrate and persist in groundwater systems over decades.

Fig. 4 shows a detailed interpretation of these dominant features. The importance of the year should not be interpreted as evidence that calendar time is itself a physical driver of PFAS contamination. Rather, this variable likely captures temporal structure embedded in the monitoring record, including changes in regulatory attention, analytical practice, reporting regimes, sampling intensity, and the evolving list of monitored analytes. Thus, its importance is best understood as predictive within the observed EPA/UCMR context, not as a direct mechanistic environmental explanation. Similarly, MRL should be interpreted as part of the measurement and reporting context rather than as a causal contamination source. Because contamination labels depend partly on analytical detectability, MRL can contribute predictive information under the same monitoring regime. However, this predictive utility should not be conflated with causal environmental influence. This distinction is important for understanding both the strengths and the limitations of the framework when applied beyond the observed analytical setting.

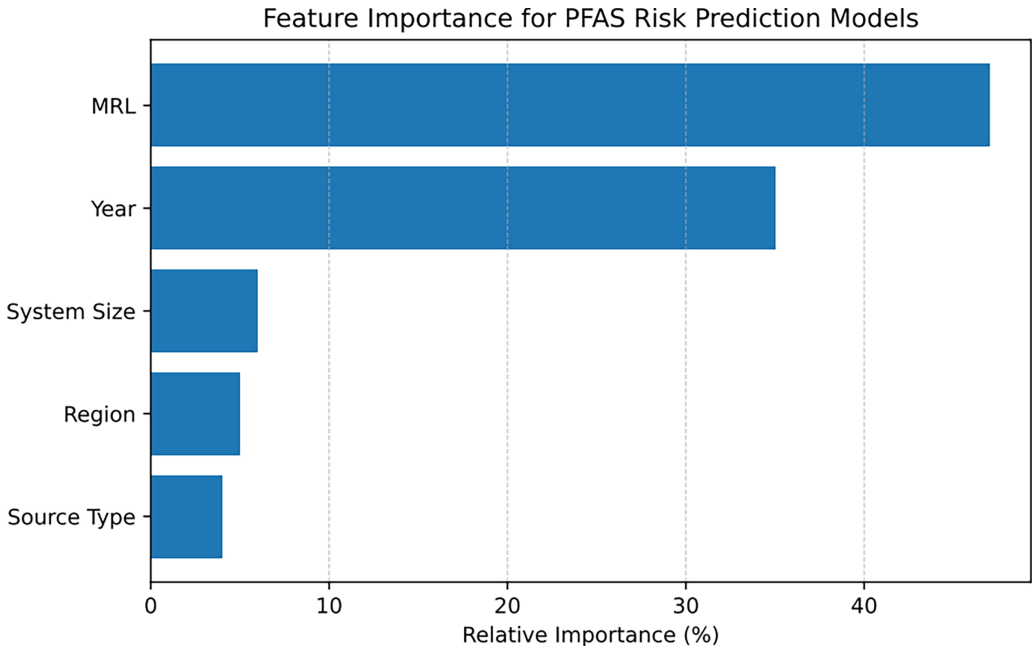

Figure 4: Top 5 feature importance distributions.

Table 4 shows the determinants of pfas detection risk. The dominance of MRL over Region suggests that the analytical detection regime contributes substantially to the observed prediction structure. This does **not** imply that geography is unimportant environmentally; rather, it indicates that, within the current dataset, measurement sensitivity exerts a strong influence on the recorded contamination label.

Table 4: Determinants of PFAS detection risk.

Rank	Feature	Importance	Interpretation
1	MRL (Detection Limit)	47%	Technological Driver: Risk is heavily dependent on analytical sensitivity. As labs detect smaller amounts (ng/L vs. µg/L), "risk" labels increase.
2	Year (Time)	35%	Regulatory Driver: Reflects the historical expansion of EPA monitoring programs and the evolving list of target analytes.

(Continued)

Table 4 (continued)

Rank	Feature	Importance	Interpretation
3	System Size	6%	Operational Driver: Larger systems serve more people and are subject to more rigorous testing schedules.
4	Region	5%	Geospatial Driver: Captures clusters of contamination, likely linked to industrial zones or aquifer characteristics.

4.6 Operational Recommendations

Finally, we translate these statistical findings into actionable advice that aligns with global assessment goals, which advocate for data-driven risk management to mitigate long-term environmental exposure. Table 5 provides a decision matrix for selecting the appropriate model based on resource constraints.

Table 5: Deployment recommendations based on resource constraints.

Regulatory Scenario	Constraint	Recommended Model	Rationale
Emergency Screening	High Budget/High Risk	Stacking Ensemble	Maximizes Recall. The priority is to find every possible contamination source, regardless of testing cost.
Routine Monitoring	Limited Budget	XGBoost (Single)	Balances speed and accuracy. Good for general surveillance where computational resources are limited.
Preliminary Audit	Low Computational Power	Random Forest	Provides a decent baseline with no need for complex hyperparameter tuning.

5 Conclusion and Future Work

This study developed an ensemble machine-learning framework for PFAS risk screening using the EPA/UCMR dataset. Voting, Weighted Averaging, and Stacking consistently outperformed the individual baseline models, confirming the value of multi-model integration for environmental risk classification. Among them, the Stacking Ensemble achieved the best overall performance, offering the most balanced precision-recall tradeoff and improving the F1-score by approximately 4% over the best single model. These findings indicate that ensemble learning can improve screening reliability without requiring additional external data.

Despite these gains, the improvement margin remained moderate, suggesting that the predictive capacity of the current chemical and monitoring variables may be approaching saturation. Therefore, the proposed framework should be regarded as an operational screening tool rather than a mechanistic forecasting model. Its applicability may also be constrained by dependence on variables associated with the monitoring regime, such as Year and MRL, which may reduce transferability across future datasets or different analytical settings. Future work should focus on incorporating physically meaningful predictors,

including source proximity, hydrogeological conditions, and treatment-infrastructure variables, to improve interpretability and generalizability. Further investigation should also examine class-imbalance handling, feature importance, and the performance tradeoffs among voting, blending, and stacking strategies. Overall, this study provides a reproducible and practically relevant framework for PFAS screening and offers a useful reference for rare-event detection in environmental monitoring systems.

Acknowledgement: This work was supported by the King Saud University, Riyadh, Saudi Arabia.

Funding Statement: This work was supported by the King Saud University, Riyadh, Saudi Arabia under the Ongoing Research funding program (ORF-2026-476).

Author Contributions: Menahil Rahman and Farhan Amin: conceptualization and writing—original draft preparation, Waqas Ishtiaq: methodology, Amerah Alabrah: software, Arif Mehmood and Rana Faraz Ahmed: validation, Iqra Khalid and Amerah Alabrah: software and writing—original draft preparation, Farhan Amin: software, investigation, and conceptualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available in the Occurrence Data from the Unregulated Contaminant Monitoring Rule, https://www.epa.gov/dwucmr/occurrence-data-unregulated-contaminant-monitoring-rule (accessed on 13 March 2026).

Ethics Approval: This study was conducted using publicly available, therefore, formal ethics approval was not required.

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

References

1. Kannan K, Corsolini S, Falandysz J, Fillmann G, Kumar KS, Loganathan BG, et al. Perfluorooctanesulfonate and related fluorochemicals in human blood from several countries. Environ Sci Technol. 2004;38(17):4489–95. doi:10.1021/es0493446.
2. Lee JC, Smaoui S, Duffill J, Marandi B, Varzakas T. Forever chemicals PFAS global impact and activities, cascading consequences of colossal systems failure: Long-term health effects, food-systems, eco-systems. Preprints. 2025.
3. Sunderland EM, Hu XC, Dassuncao C, Tokranov AK, Wagner CC, Allen JG. A review of the pathways of human exposure to poly- and perfluoroalkyl substances (PFASs) and present understanding of health effects. J Expo Sci Environ Epidemiol. 2019;29(2):131–47. doi:10.1038/s41370-018-0094-1.
4. Hu XC, Andrews DQ, Lindstrom AB, Bruton TA, Schaider LA, Grandjean P, et al. Detection of poly- and perfluoroalkyl substances (PFASs) in US drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants. Environ Sci Technol Lett. 2016;3(10):344–50. doi:10.1021/acs.estlett.6b00260.
5. Reinikainen J, Bouhoulle E, Sorvari J. Inconsistencies in the EU regulatory risk assessment of PFAS call for readjustment. Environ Int. 2024;186(1):108614. doi:10.1016/j.envint.2024.108614.
6. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
7. Friedman JH. Greedy function approximation: a Gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
8. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: The Association for Computing Machinery (ACM); 2016.
9. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient Gradient boosting decision tree. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA.
10. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology. 2007;88(11):2783–92. doi:10.1890/07-0539.1.

11. Olawade DB, Ijiwade JO, Fapohunda O, Ige AO, Olajoyetan DO, Wada OZ. Predictive modeling of PFAS behavior and degradation in novel treatment scenarios: a review. Process Saf Environ Prot. 2025;196(15):106869. doi:10.1016/j.psep.2025.106869.
12. Islam MJ, Salekin SU, Anzum A, Zaman N, Khan AA, Sarkar D, et al. Machine learning-driven water quality index prediction: enhancing accuracy with Gradient boosting and explainable AI for sustainable water monitoring. Appl Agric Sci. 2024;2(1):1–14.
13. Doorn N. Artificial intelligence in the water domain: opportunities for responsible use. Sci Total Environ. 2021;755(9):142561. doi:10.1016/j.scitotenv.2020.142561.
14. Elumalai V, Brindha K, Sithole B, Lakshmanan E. Spatial interpolation methods and geostatistics for mapping groundwater contamination in a coastal area. Environ Sci Pollut Res. 2017;24(12):11601–17. doi:10.1007/s11356-017-8681-6.
15. El Ouali A, Bayhan K, Mouhoumed RM, Spor P, Atan CS, Başakın EE, et al. Performance of tree-based ensemble techniques in predicting groundwater quality for irrigation purposes. Environ Earth Sci. 2025;84(16):474. doi:10.1007/s12665-025-12469-w.
16. Misstear B, Vargas CR. A global perspective on assessing groundwater quality. Hydrogeol J. 2023;31(1):11–4. doi:10.1007/s10040-022-02461-0.
17. Liu K. Environment risk assessment of healthcare waste using ensemble learning technique-based EFMEA. Int J Rel Qual Saf Eng. 2024;31(6):2450022. doi:10.1142/s0218539324500220.
18. Cha GW, Moon HJ, Kim YC. Comparison of random forest and Gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. Int J Environ Res Public Health. 2021;18(16):8530. doi:10.3390/ijerph18168530.
19. Zhou ZH. Ensemble methods: foundations and algorithms. Boca Raton, FL, USA: CRC Press; 2025.
20. Kuncheva LI. Combining pattern classifiers. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2014. doi:10.1002/9781118914564.
21. Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag. 2006;6(3):21–45. doi:10.1109/MCAS.2006.1688199.
22. Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Berlin/Heidelberg, Germany: Springer; 2000. p. 1–15. doi:10.1007/3-540-45014-9_1.
23. Bishop C, Nasrabadi M. Pattern recognition and machine learning. Berlin/Heidelberg, Germany: Springer; 2006.
24. Vapnik V, Izmailov R. Rethinking statistical learning theory: learning using statistical invariants. Mach Learn. 2019;108(3):381–423. doi:10.1007/s10994-018-5742-0.
25. Gacu JG, Monjardin CEF, Mangulabnan RGT, Pugat GCE, Solmerin JG. Artificial intelligence (AI) in surface water management: a comprehensive review of methods, applications, and challenges. Water. 2025;17(11):1707. doi:10.3390/w17111707.
26. Masud Rana SM, Boccelli DL. Contaminant spread forecasting and confirmatory sampling location identification in a water-distribution system. J Water Resour Plann Manage. 2016;142(12):04016059. doi:10.1061/(asce)wr.1943-5452.0000704.
27. Román JJ, Ramos LT, Ketbi AA, Dhaheri SA, Rivas-Echeverría F. Deep learning for environmental monitoring and conservation: applications, approaches, challenges, and future perspectives. TechRxiv:176463786.63610597. 2025.
28. Ngwenya B, Paepae T, Bokoro PN. Monitoring ambient water quality using machine learning and IoT: a review and recommendations for advancing SDG indicator 6.3.2. J Water Process Eng. 2025;73(2):107664. doi:10.1016/j.jwpe.2025.107664.
29. Job C. Responding to EPA's PFAS strategic roadmap. Groundw Monit Remediat. 2024;44(4):21–6. doi:10.1111/gwmr.12686.
30. Dams R, Ameduri B. Essential per- and polyfluoroalkyl substances (PFAS) in our society of the future. Molecules. 2025;30(15):3220. doi:10.3390/molecules30153220.
31. Post GB. Recent US state and federal drinking water guidelines for per- and polyfluoroalkyl substances. Environ Toxicol Chem. 2021;40(3):550–63. doi:10.1002/etc.4863.

32. Shahbaz M, Gao C, Zhai L, Shahzad F, Khan I. Environmental air pollution management system: predicting user adoption behavior of big data analytics. Technol Soc. 2021;64:101473. doi:10.1016/j.techsoc.2020.101473.
33. Hong J, Kang H, Hong T. Oversampling-based prediction of environmental complaints related to construction projects with imbalanced empirical-data learning. Renew Sustain Energy Rev. 2020;134(1):110402. doi:10.1016/j.rser.2020.110402.
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Jair. 2002;16:321–57. doi:10.1613/jair.953.
35. Brownlee J. Imbalanced classification with python: better metrics, balance skewed classes, cost-sensitive learning. Melbourne, VIC, Australia: Machine Learning Mastery; 2020.
36. Kombo Mpindou GOM, Escuder Bueno I, Chordà Ramón E. Risk analysis methods of water supply systems: comprehensive review from source to tap. Appl Water Sci. 2022;12(4):56. doi:10.1007/s13201-022-01586-7.
37. Sharma S, Mishra D, Yadav A, Gami B, Madhan ES. An Intelligent, low-cost water quality monitoring system with on-device machine learning and cloud integration. Sci Rep. 2026;16(1):1600. doi:10.1038/s41598-026-37287-3.
38. Molinari DA. Spatiotemporal modelling of groundwater contaminants [dissertation]. Glasgow, UK: University of Glasgow; 2014.
39. Udu AG, Salman MT, Ghalati MK, Lecchini-Visintini A, Siddle DR, Dong H. Emerging SMOTE and GAN variants for data augmentation in imbalance machine learning tasks: a review. IEEE Access. 2025;13(6):113838–53. doi:10.1109/access.2025.3584532.
40. Cascajares M, Alcayde A, Salmerón-Manzano E, Manzano-Agugliaro F. The bibliometric literature on Scopus and WoS: the medicine and environmental sciences categories as case of study. Int J Environ Res Public Health. 2021;18(11):5851. doi:10.3390/ijerph18115851.
41. Liu J, Sun L, Tian Z, Ye Q, Wu S, Zhang S. Nature-based solutions for urban water management. Front Environ Sci. 2023;11:1228154. doi:10.3389/fenvs.2023.1228154.
42. Division HAM. Guidance on PFAS exposure, testing, and clinical follow-up. Washington, DC, USA: National Academies Press; 2022. doi:10.17226/26156.
43. EU Chemicals Agency. [cited 2026 Mar 4]. Available from: https://www.euronews.com/my-europe/2026/03/26/eu-chemicals-agency-backs-forever-chemicals-ban-with-final-decision-to-the-commission?utm_source=chatgpt.com.
44. USGS No. 1490. Integrated science for the study of perfluoroalkyl and polyfluoroalkyl substances (PFAS) in the environment—A strategic science vision for the US Geological Survey. Reston, VA, USA: US Geological Survey; 2021.
45. Ramdani F. Data science: foundations and hands-on experience: handling economic, spatial, and multidimensional data with R. Singapore: Springer Nature; 2025. doi:10.1007/978-981-96-4683-8.