

Data driven prediction of uranium in groundwater for Environmental Radioactivity Surveillance

Padma Savitri P.

ppsavitri@barc.gov.in

BARC

Utkarsh Bharadwaj

BARC

Ramesh B

BARC

Sudhakar J

BARC

Balaram kumar R

BARC

Manoj Warriar

BARC

A.D.P. Rao

Andhra University

Research Article

Keywords: Groundwater, Uranium, Machine learning, Prediction, Hydrogeochemistry

Posted Date: June 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9854694/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Environmental Geochemistry and Health on September 28th, 2026. See the published version at <https://doi.org/10.1007/s10653-026-03510-3>.

**Data driven prediction of uranium in groundwater for
Environmental Radioactivity Surveillance**

- **Corresponding author:** P.Padma Savitri, Environmental Monitoring & Assessment Division, BARC, Visakhapatnam, Andhra Pradesh, India-531011. Email: ppsavitri@barc.gov.in, psavitrip.pps@gmail.com; Phone: +91-891-2838033, 9550369981

Co-authors in the following order:

- Utkarsh Bharadwaj, Computational Analysis Division, BARC, Visakhapatnam, Andhra Pradesh, India-531011 email : butkarsh@barc.gov.in
- B.Ramesh, Environmental Monitoring & Assessment Division, BARC, Visakhapatnam, Andhra Pradesh, India-531011 email : rameshb@barc.gov.in
- J.Sudhakar, Environmental Monitoring & Assessment Division, BARC, Visakhapatnam, Andhra Pradesh, India-531011 email: jsudhakar@barc.gov.in
- R.Balaram kumar, Environmental Monitoring & Assessment Division, BARC, Visakhapatnam, Andhra Pradesh, India-531011 email: balaramkr@barc.gov.in
- Manoj Warriar, Computational Analysis Division, BARC, Visakhapatnam, Andhra Pradesh, India-531011 email : manoj.warrier@gmail.com
- A.D.P.Rao, Nuclear Physics Department, Andhra University, Visakhapatnam, Andhra Pradesh, India-530003 email: adprao18@gmail.com

Data driven prediction of uranium in groundwater for Environmental Radioactivity Surveillance

Abstract

Uranium occurrence in groundwater is governed by complex hydrogeochemical processes, yet routine monitoring remains analytically demanding. Developed a machine learning based framework for predicting uranium concentrations in groundwater using routinely measured physicochemical parameters that reflect the underlying hydrogeochemical controls. Around 1300 groundwater samples from a coastal region of southeastern India were analyzed, for uranium concentrations with conventional water quality parameters. A threshold driven binary classification approach was implemented using three key limits ($2\mu\text{gL}^{-1}$, $15\mu\text{gL}^{-1}$, and $30\mu\text{gL}^{-1}$), reflecting precautionary, guideline and regulatory based limits. Advanced ensemble learning algorithms were evaluated under pronounced class imbalance conditions. CatBoost demonstrated superior predictive capability, achieving an F1-score of 80.8%, while Light GBM yielded an F1-score of 66.6% at the $2\mu\text{gL}^{-1}$ and $15\mu\text{gL}^{-1}$ thresholds. An anomaly detection model achieved perfect recall with a 1.12% false positive rate, reducing laboratory screening by 98.7%. Feature importance analysis identified total dissolved solids (TDS), hardness, chloride, and sulphate as dominant predictors. These parameters reflect key hydrogeochemical processes, including mineral dissolution, salinity evolution, and carbonate complexation, which govern uranium mobility in groundwater systems. SHAP-based interpretability analysis confirmed the mechanistic relevance of TDS, hardness, and sulphate in driving model predictions. Spatial cross-validation demonstrated robust generalization across distinct hydrogeochemical zones. This study represents the first attempt to apply machine learning for the predictive assessment of uranium in groundwater in a coastal region of India using measurable physicochemical parameters. The approach provides a cost-effective tool for environmental monitoring, early warning, and sustainable groundwater management, with important implications for public health and regulatory decision-making.

Keywords: Groundwater, Uranium, Machine learning, Prediction, Hydrogeochemistry

Article Highlights

- *Machine learning framework developed to predict uranium concentration in groundwater using routine water quality parameters.*
- *Total Dissolved Solids (TDS), hardness, chloride, and sulphate are key predictive variables.*
- *The framework enables cost-effective screening, early warning, and optimized groundwater monitoring strategies.*

Data driven prediction of uranium in groundwater for Environmental Radioactivity Surveillance

Introduction

Water is a critical natural resource, and groundwater quality is influenced by both natural and anthropogenic factors. Hydrogeochemical conditions and geogenic processes strongly influence the occurrence of radionuclides in groundwater systems (Alinne Marianne Martins Araújo et al., 2026). Naturally occurring uranium is highly soluble and mobile, leading to elevated concentrations in aquifers compared to other radionuclides (Balaram et al. 2022). In India, uranium levels in groundwater range from 0.2 to 4918 μgL^{-1} (mean 6.8 μgL^{-1}) (Sahoo et al. 2021), posing chemical and radiological risks, particularly to kidneys and lungs (UNSCEAR 2016). Despite regulatory limits (BIS 2021; AERB 2004; WHO 2012), many rural communities consume groundwater without testing, increasing exposure risk. Uranium mobility is governed by complex interactions among redox conditions, pH, mineral content, and competing ions, controlling its speciation and solubility (Sahoo et al. 2021). Uranium concentration in groundwater, its mobility and mineral precipitation, is strongly controlled by physicochemical conditions such as pH, redox potential, and overall water chemistry (Nikitenkov A.N. et al. 2025). Conventional detection methods to analyse uranium routinely, while accurate, are costly and impractical for large-scale monitoring, underscoring the need for predictive, cost effective approaches.

Machine learning (ML) has emerged as a powerful data-driven approach for capturing complex nonlinear relationships among hydrogeochemical variables. The effectiveness of data-driven approaches in hydrogeochemical studies is increasingly supported by recent literature (Käßner A et al., 2022). Their study demonstrated that machine learning models, when coupled with multiple environmental variables, can successfully capture spatial variability in geochemical signatures and improve predictive accuracy, particularly in geologically complex regions. In parallel, ML has been widely applied to model common groundwater contaminants such as fluoride, nitrate, and chloride (Chen et al., 2018; Knoll et al., 2019; Nafouanti et al., 2021; Gupta et al., 2022; Podgorski et al., 2022; Yang et al., 2023; Anjali et al., 2024; Durga Prasad Panday et al., 2024; Karunanidhi et al., 2025).

Earlier studies have predominantly relied on statistical and regression-based methods, including multiple linear regression, logistic regression, and Principal Component Analysis (PCA),(Ayotte et al., 2011;Kumar et al. 2020) While these methods provide interpretable coefficients and are computationally efficient, they suffer from critical limitations: (i) assumption of linear relationships between predictors and uranium concentration, which fails to capture the complex non-linear geochemical processes governing uranium mobility (e.g.,pH-dependent solubility, redox-controlled speciation); (ii) poor performance under class imbalance, as standard logistic regression assigns equal misclassification costs to both classes; and (iii) inability to model feature interactions without explicit polynomial or interaction terms, which increases dimensionality and the risk of over fitting.

The latest study (Anjali Kerketta et al., 2026) demonstrated the application of machine learning for predicting elevated uranium concentrations in groundwater, employing Random Forest–based spatial modelling in Punjab, India, using predictor variables related to meteorological, topographical, geological, soil, and other relevant parameters.

However, the application of machine learning for predicting uranium concentrations in groundwater using routinely monitored physicochemical parameters is still emerging, with relatively limited studies reported to date.

Addressing these limitations, the present study developed a novel ML framework for prediction of uranium in groundwater, integrating (i) routine physicochemical parameters, (ii) multi-threshold labeling aligned with WHO and national standards, and (iii) uncertainty quantification to assess prediction reliability. By combining these elements, the framework offers a first of its kind, operationally deployable tool for scalable, reliable, and actionable uranium screening, bridging a critical gap between hydrogeochemistry, predictive modeling, and practical groundwater management.

Machine learning methods, particularly ensemble tree-based algorithms, overcome these limitations by: (a) capturing non-linear relationships through recursive partitioning, (b) handling class imbalance via class weighting and specialized loss functions, and (c) automatically learning complex feature interactions without extensive feature engineering. Ratio-based features have been demonstrated to enhance uranium prediction models by capturing hydrogeochemical relationships that influence uranium mobility, such as the interplay between major ions affecting carbonate complexation and pH buffering capacity (Nolan and Weber, 2015; Coyte et al., 2018).

Key Methodological Contributions

This study presents a novel application of machine learning modeling for the prediction of uranium concentration in groundwater along the south eastern coast of India. To the best of our knowledge this is the first attempt in the study region to classify groundwater uranium levels using machine learning models based exclusively on independent hydrogeochemical parameters, without incorporating uranium as an input feature. Unlike conventional statistical approaches, the proposed framework enables robust, non-linear prediction of uranium distribution and identifies key radiochemical controls influencing groundwater uranium levels. The findings provide a new data driven tool for groundwater quality assessment and contribute directly to public health assessment and management in uranium prone coastal aquifers. Multi threshold classification frameworks with thresholds 2, 15, and 30 $\mu\text{g L}^{-1}$, corresponding to the historical evolution of WHO drinking water guidelines (1998, 2004, 2011), were employed. This approach enables hierarchical risk stratification across different regulatory contexts, supports early warning detection before regulatory limits are approached, provides operational flexibility for different monitoring objectives and links scientific understanding of uranium toxicity to classification strategy. We have applied one class anomaly detection (Isolation Forest) for uranium exceedance detection at extreme class imbalance (99.85% negative class). This methodological innovation addresses the failure of traditional binary classifiers at extreme imbalance, provides a conservative screening approach prioritizing recall (zero false negatives) and offers a practical utility for regulatory compliance monitoring where missing exceedances is unacceptable.

This study integrates Bootstrap confidence intervals (95%) for all performance metrics, Spatial cross-validation to evaluate geographic transferability, Uncertainty bands on calibration plots and PR curves and Statistical significance testing (McNamara's test) for model comparison. This rigorous validation framework exceeds standard practices in groundwater quality ML studies. We provide SHAP-based feature importance using only measured chemical parameters (no spatial/temporal features), Decision threshold optimization with operational justification, deployment guidelines for different monitoring scenarios (pre-operational, routine, resource-constrained, regulatory compliance) and Cost-benefit analysis demonstrating practical utility. These contributions advance the uranium prediction studies by providing a complete, operationally deployable framework for groundwater monitoring with robust uncertainty quantification and regulatory alignment.

Materials and methods

The study was carried out in a $\sim 1,500 \text{ km}^2$ region surrounding the Bhabha Atomic Research Centre (BARC), Visakhapatnam, facility located near Jogannapalem village ($17^\circ 31' 23.9''\text{N}$, $83^\circ 02' 09.9''\text{E}$) located along the southeastern coast of India. The region was selected considering the upcoming nuclear facilities, including NPCIL's proposed 1000 MWe reactors at Kovvada and BARC research facilities at Atchutapuram. Groundwater samples were collected from eight landward sectors extending up to 50 km from the BARC site between 2016 and 2025. Groundwater samples were collected following standard BIS protocols using pre-cleaned polyethylene bottles and analyzed for uranium and major physicochemical parameters including pH, EC, TDS, salinity, hardness, fluoride, chloride, nitrate, sulfate, and phosphate. Uranium concentration was determined using an LED-based UV fluorimeter (Model UA2, Quantalase Enterprises Pvt. Ltd., India), a sensitive and cost-effective technique widely employed for trace uranium estimation in water samples (Sahoo et al. 2021). Detailed information on the study area, sampling design, and spatial distribution has been reported

(P.PadmaSavitri et al. 2026). The compiled dataset was subsequently utilized for machine learning modeling and predictive analysis.

In the present study, uranium concentration was treated solely as the target variable for classification. The predictor set comprised only independent groundwater physicochemical parameters and uranium was intentionally excluded from the input feature space to prevent target leakage.

Data preprocessing, labeling and, Quality control

Data preprocessing involved systematic cleaning to ensure data consistency and reliability, including removal of duplicates, correction of entry errors, unit standardization, and handling of missing values. Potential outliers were identified using the Isolation Forest algorithm and the interquartile range (IQR) method and were verified against field and laboratory records to distinguish genuine extreme concentrations from errors. Samples with unresolved inconsistencies or incomplete multi-parameter data were excluded, resulting in a final dataset of 1,295 groundwater samples. Uranium concentrations were labeled using health-based guideline thresholds recommended by the World Health Organization (WHO). Binary classification was performed using three thresholds (2, 15, and 30 μgL^{-1}), corresponding to successive WHO guideline revisions, with 30 μgL^{-1} representing the current provisional value.

The multi-threshold labeling approach is justified as it enables comparison across historical and current regulatory standards, supports early detection of elevated uranium concentrations prior to exceedance, and reduces the impact of extreme class imbalance on model training and evaluation. From an operational perspective, this strategy allows targeted monitoring and prioritization of locations requiring increased sampling frequency, thereby optimizing laboratory resources while ensuring compliance with national and international regulatory frameworks.

Feature Engineering and Selection

Feature engineering focused on physicochemical parameters directly associated with uranium geochemistry to ensure interpretability and practical applicability. Spatial and temporal variables were excluded to avoid over fitting and to enhance generalizability across sampling periods (Dougherty et al. 1995). The input features included routinely measured water-quality parameters (pH, EC, TDS, hardness, salinity, turbidity, chloride, sulfate, nitrate, phosphate, and fluoride), along with engineered ratio-based and interaction features (e.g., TDS/Hardness, Chloride/Sulfate, EC/TDS, TDS $\times$ Hardness). This approach ensures suitability for deployment in resource limited monitoring settings. (Guyon I et al.2003)

Exploratory Data Analysis: Dimensionality Reduction

Dimensionality reduction was used solely for exploratory analysis to visualize multivariate relationships between physicochemical parameters and uranium concentration. We applied t-distributed Stochastic Neighbor Embedding (t-SNE), a nonlinear method effective in preserving local structure and revealing clustering patterns in high-dimensional data (Van der Maaten, Hinton G, 2008). The analysis explored potential nonlinear associations between uranium levels and key variables (e.g., TDS, hardness, chloride, and EC), with visualizations indicating distinct groupings across concentration ranges that support the feasibility of classification. Although t-SNE lacks global interpretability and is sensitive to parameter initialization, it was not used for model training; rather, it served to inform hypothesis generation and feature relevance during exploratory data analysis. It was not used for dimensionality reduction in the final classification models; all measured physicochemical parameters were retained as input features to preserve interpretability and maintain consistency with routine monitoring protocols.

244 Spatial Cross Validation implementation

245

246 To account for potential spatial correlations in groundwater samples and provide rigorous model
247 validation for spatial generalizability, we implemented spatial cross-validation alongside traditional k-
248 fold cross-validation. We develop a spatial cross-validation framework that partitions samples into
249 geographically coherent blocks prior to model training. Spatial blocks were created using K-means
250 clustering on sample coordinates (latitude and longitude), ensuring geographic separation between
251 training and test sets in each fold. This blocking strategy mimics real-world prediction scenarios
252 where models must generalize to new, spatially distinct locations not represented in the training data.
253 The cross-validation procedure followed a standard k-fold protocol ($k=3$ for model evaluation), with
254 each iteration holding out spatial blocks for testing while training on the remaining blocks.

255

256 Classification Algorithms

257

258 A suite of supervised learning algorithms was evaluated to classify groundwater samples based on
259 predefined uranium concentration thresholds. Model performance was rigorously compared across
260 multiple validation schemes and evaluation metrics. The primary focus was on ensemble tree based
261 methods such as Random Forest (Belgiu et al. 2016; Breiman 2001), XGBoost (Chen et al. 2016),
262 LightGBM, (Ke G et al. 2017) and CatBoost (Prokhorenkova et al. 2018) due to their proven ability to
263 capture complex nonlinear relationships and high-order feature interactions, which are common in
264 hydrogeochemical datasets. In addition to these advanced ensemble models, we included interpretable
265 and margin based baseline classifiers: a Support Vector Machine (SVM) (Cortes et al. 1995) with a
266 radial basis function (RBF) kernel and a single Decision Tree. The latter served as a transparent
267 benchmark to assess the plausibility and directional consistency of learned feature relationships.
268 Given the increasing class imbalance at higher uranium thresholds, the modeling strategy was adapted
269 accordingly.

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

- At the $2 \mu\text{gL}^{-1}$ threshold (relatively balanced classes), class weighting was applied.
- At $15 \mu\text{gL}^{-1}$ (moderate imbalance), we combined class weighting (e.g., `scale_pos_weight`) with decision threshold optimization and tested synthetic minority oversampling (SMOTE) and hybrid SMOTE-Tomek methods, where they improved cross-validation performance.
- At $30 \mu\text{gL}^{-1}$ (extreme imbalance, prevalence $< 2\%$), supervised classifiers were supplemented with one-class anomaly detection methods specifically Isolation Forest and One Class SVM, to serve as conservative screening tools prioritizing high recall, ensuring no high-concentration samples were missed.

Each ensemble method offers distinct advantages for environmental data, as given below.

- Random Forest aggregates predictions from multiple decision trees trained on bootstrap samples, providing robustness to noise, outliers, and scale heterogeneity among physicochemical covariates.
- XGBoost employs gradient boosting with second-order optimization, enhancing convergence and performance, particularly when input features are unevenly distributed across spatial or geological domains.
- LightGBM uses histogram based binning and leaf wise tree growth, significantly reducing memory usage and training time while maintaining high accuracy, making it well-suited for large tabular datasets like ours.
- CatBoost implements ordered boosting and intrinsic handling of categorical variables. Although most predictors in this study are continuous, CatBoost's built-in regularization helps mitigate over fitting, especially in smaller or stratified subsets of the data.

The SVM with RBF kernel was included as a strong nonlinear baseline in high-dimensional spaces, contingent on proper feature scaling. It provided a useful contrast to tree based models in terms of decision boundary geometry and sensitivity to feature distributions. To ensure optimal performance, hyper parameters for all models were tuned using Bayesian optimization via Optuna, with up to 500 trials per model (Akiba T et al. 2019; Bergstra et al. 2012). For tree-based ensembles, key parameters included learning rate, maximum tree depth or number of leaves, minimum child/leaf sample size,

297 subsampling and column sub sampling ratios, and L1/L2 regularization strengths. For SVM, the
298 penalty parameter C and kernel coefficient γ were optimized. To address class imbalance, we
299 employed a combination of strategies: class weighting, decision threshold tuning to reflect asymmetric
300 misclassification costs (e.g., prioritizing recall for high-risk classes), and light resampling
301 (SMOTE/SMOTE-Tomek) only when it yielded measurable improvements in validation performance.
302

303 **Model Evaluation**

304
305 To comprehensively assess model performance and generalizability, we implemented a dual
306 evaluation framework that measures both predictive accuracy and geographic transferability. First,
307 stratified five-fold cross-validation (CV) was employed to ensure balanced class representation across
308 all folds, thereby preserving the original class distribution and reducing variance associated with a
309 single train-test split (Roberts et al. 2017; Kohavi 1995; Efron B 1979). This approach provides a
310 robust estimate of model performance on randomly sampled data. Second, to evaluate spatial
311 generalization, a critical requirement for environmental models, we implemented spatial cross-
312 validation. In this scheme, entire geographic locations or spatial blocks were held out during training
313 and used exclusively for testing. This strategy mitigates optimistic bias arising from spatial
314 autocorrelation, where nearby samples are more similar than distant ones, and provides a more
315 realistic assessment of a model's ability to predict uranium concentrations in unsampled regions
316 (Roberts et al. 2017).

317 K-Means clustering strategy on the sample coordinates (Latitude/Longitude) to partition the
318 study area into 5 distinct spatial blocks (`n_splits=5`) was employed. This approach ensures that
319 samples from the same geographic cluster are kept together in either the training or testing set,
320 preventing spatial autocorrelation leakage. The random seed was set to 42 for reproducibility.
321 All preprocessing steps, including missing value imputation, feature scaling (for SVM), class
322 weighting, and probability calibration, were applied strictly within the training partition of each fold.
323 Hyper parameter tuning was nested within the cross-validation loop using Bayesian optimization
324 (Optuna), ensuring that no information from the validation or test sets influenced model training or
325 selection (Akiba et al. 2019; Bergstra et al. 2012). This rigorous protocol prevents data leakage and
326 ensures unbiased performance estimates. We report a comprehensive set of evaluation metrics tailored
327 to different operational and regulatory needs (Chinchor N 1992). These include:

- 328 • Accuracy, Precision, Recall (Sensitivity), and the F1-score (harmonic mean of precision and
329 recall).
- 330 • Area under the ROC curve (AU-ROC), which evaluates overall discriminative ability.

331 Area under the Precision Recall curve (PR-AUC), Balanced Accuracy, and the Matthews Correlation
332 Coefficient (MCC) metrics are particularly informative under class imbalance (Saito & Rehmsmeier
333 2015).

334 The complete workflow from data collection to decision support is illustrated in Fig. 1. Given the
335 increasing rarity of positive cases at higher thresholds (especially $\geq 30 \mu\text{gL}^{-1}$), PR-AUC and MCC
336 were prioritized as primary metrics for evaluating rare event detection performance, as they are less
337 sensitive to class imbalance than AU-ROC. Performance was reported across both random five-fold
338 Cross Validation and spatial Cross Validation to contrast model behavior under idealized and real-
339 world deployment conditions. We present F1-scores, ROC curves, and PR curves to support detailed
340 analysis of probability calibration and classification thresholds. Finally, SHAP (SHapley Additive
341 exPlanations) analysis, based solely on measured physicochemical features, was used for model
342 interpretation. It revealed Total Dissolved Solids (TDS), Hardness, and Chloride as the top predictors
343 at the $2 \mu\text{gL}^{-1}$ threshold. SHAP dependence plots indicate a positive association between uranium
344 exceedance probability and elevated levels of TDS, hardness, and chloride trends that align well with
345 established geochemical mechanisms governing uranium mobility in groundwater systems.
346

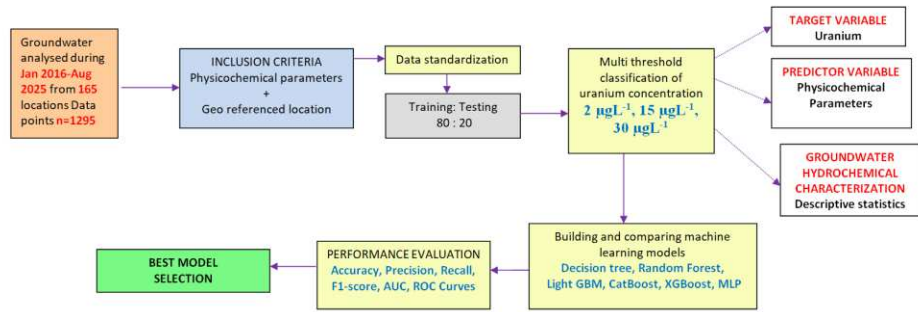


Fig. 1 Flow diagram demonstrating the methodology applied using machine learning models to predict uranium concentration

Results

A total of approximately 1,300 groundwater samples were collected and analyzed during 2016 to 2025 from around 165 distinct locations across the study region. During the study period, no operational radiological or nuclear facilities were present in or near the selected area, ensuring that observed uranium levels reflect natural geochemical processes rather than anthropogenic contamination. All measured physicochemical parameters, including pH, electrical conductivity, total dissolved solids (TDS), and major ions, had average values within the permissible limits specified by the Bureau of Indian Standards (BIS) for drinking water quality (BIS, 2012). Furthermore, uranium concentrations in all samples were below the Indian regulatory limit of $60 \mu\text{g/L}$ set by the Atomic Energy Regulatory Board (AERB 2004), indicating no immediate regulatory exceedance across the region. This study introduces several methodological advances in the machine learning based prediction and classification of uranium in groundwater, beginning with a comprehensive exploratory data analysis.

Exploratory Data Analysis

The groundwater dataset exhibited considerable variability across all measured physicochemical parameters. Stastical summary of physicochemical parameters is presented in Table 1 (P.Padma Savitri et al. 2026). All variables showed significant deviation from normality, as confirmed by the Shapiro-Wilk test ($p < 0.0001$). Pronounced skewness was observed in turbidity (4.7), chloride (3.4), Hardness (3.7) and total dissolved solids (TDS) (3.97). These non-normal distributions necessitated the use of non-parametric statistical methods and appropriate data transformations (e.g., logarithmic scaling) prior to modeling to ensure robust algorithm performance.

Uranium concentrations ranged from $0.1 \mu\text{g/L}$ to $38 \mu\text{g/L}$, with a mean of $3.9 \pm 4.4 \mu\text{g/L}$ and a median of $2.6 \mu\text{g/L}$, indicating a right skewed distribution (Fig. 2). This pattern is characteristic of natural uranium occurrence in aquifers, where the majority of samples have low concentrations, while a small fraction exhibit elevated levels. Summary table showing the distribution of samples across the specified uranium concentration thresholds is presented in Table 2. The frequency distribution of uranium concentrations (Fig. 2) highlights the proportion of samples exceeding key World Health Organization guideline thresholds. This distribution reflects the natural geochemical variability of uranium and strongly supports the adoption of a multi-threshold classification framework, enabling early detection, risk stratification, and alignment with evolving regulatory standards. To qualitatively assess the relationship between uranium concentration and key physicochemical parameters, t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten, Hinton 2008) was applied to project the six-dimensional feature space (comprising the parameters listed in Table 1) into a two-dimensional representation.

390 **Table 1 Comprehensive Statistical summary of water quality parameters (2016-2025)**
 391

Parameter	Range	Mean±SD	Median	IQR	CV	Skewness	Kurtosis
pH	5.4-8.9	7.43±0.49	7.4	0.66	0.067	0.06	0.306
EC (μScm^{-1})	94.9-13630	1784±1544	1322	1254	0.861	2.55	9.38
TDS (mgL^{-1})	66.7-15300	1187±1133	860	774	0.952	3.97	28.37
Hardness(mgL^{-1})	15-2086	251±192	210	184	0.739	3.73	22.5
Chloride (mgL^{-1})	1.4-3243	264±339	57	96	1.791	3.39	15.92

392 IQR: Interquartile range; CV: Coefficient of variation, SD: Standard Deviation
 393

Table 2: Distribution of Uranium concentration

Uranium concentration (μgL^{-1})	No. of Samples	Percentage (%)
<2	545	42.08%
2-15	705	54.44%
15-30	43	3.32%
≥ 30	2	0.15%
Total	1295	100%

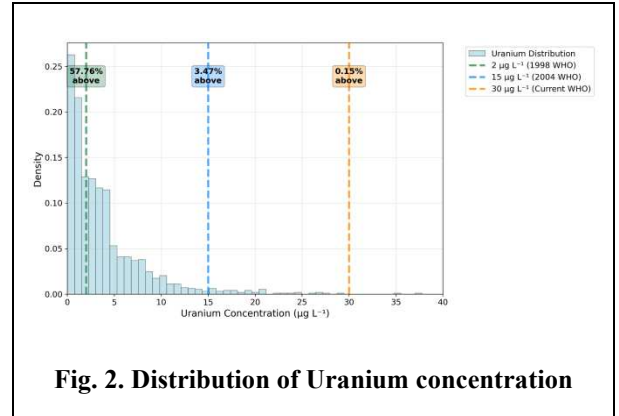


Fig. 2. Distribution of Uranium concentration

394
 395 Using the same dataset, Pearson correlation analysis was carried out to identify the factors influencing
 396 uranium mobility and groundwater quality variations across the study area, and the detailed results
 397 have been reported (P.Padma Savitri et al). None of the parameters have correlation coefficients
 398 greater than 0.85, indicating there is no severe multicollinearity among the predictor variables, which
 399 is desirable for machine learning and regression models. Tree-based gradient boosting models like
 400 XGBoost, LightGBM, and CatBoost which show the best performance in our study are inherently
 401 robust to multicollinearity. Unlike linear models, they make splits based on most information gain
 402 rather than relying on linear coefficients. Consequently, correlated features do not destabilize the
 403 model and we retain all the features for training our models. In addition feature contributions are
 404 assessed using SHAP, enabling transparent interpretation of variable importance, non linear effects,
 405 and interactions.
 406

t-SNE Analysis: Chemical Parameters vs Uranium Concentration
WHO Three-Threshold Classification

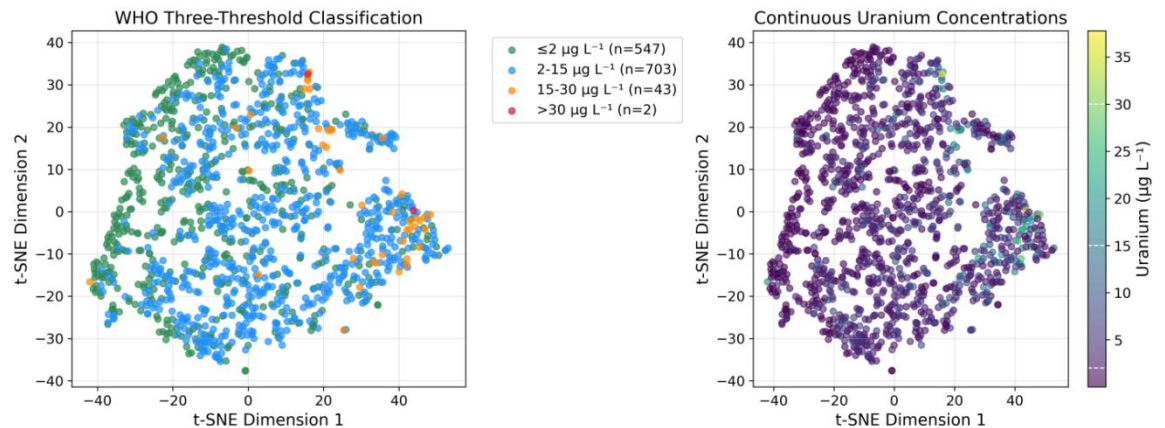
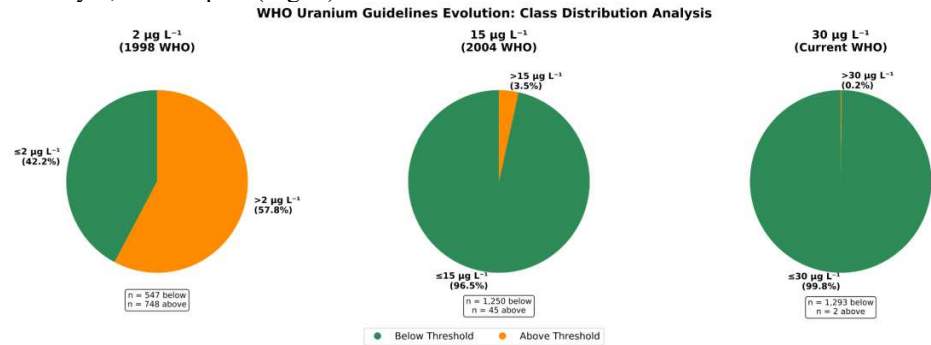


Fig. 3. t-SNE visualization of chemical parameters vs uranium concentration

409 This nonlinear dimensionality reduction technique preserves local data structures, allowing
410 visualization of potential clustering patterns in the high-dimensional space. Fig.3 presents the 2D t-
411 SNE embedding, with points colored by (a) uranium concentration categories (<2, 2–15, and 15–
412 30 μgL^{-1}), and (b) actual uranium concentration values. Each point represents a groundwater sample,
413 and its color reflects uranium levels to visually evaluate the collective predictive capacity of the six
414 parameters. The visualization reveals distinct clustering patterns, with low-uranium samples
415 predominantly grouped on the left and higher concentration samples distributed towards the right.
416 Although a sharp, continuous gradient is not evident, a coarse trend is observable along the x-axis,
417 suggesting increasing uranium levels with higher values in the reduced feature space. These patterns
418 indicate that the selected physicochemical parameters collectively capture meaningful variance related
419 to uranium concentration, supporting their utility as predictors in machine learning models. The
420 presence of separable clusters, particularly between low and high concentration groups, suggests that
421 nonlinear classification methods are well-suited for this task.

422
423 **Multi-threshold classification results**

424
425 A three tiered classification framework was implemented to evaluate model performance across
426 different regulatory and scientific thresholds: 2 μgL^{-1} , 15 μgL^{-1} , and 30 μgL^{-1} providing a scientifically
427 grounded framework for hierarchical risk assessment. This approach reveals distinct challenges and
428 opportunities in uranium prediction, reflecting varying degrees of class imbalance across the dataset of
429 approximately 1,295 samples (Fig. 4).



430
431

432 **Fig. 4. Uranium classification thresholds comparison**

433
434 At the 2 μgL^{-1} threshold, the class distribution was relatively balanced (approximately 57.8% above,
435 42.2% below), enabling robust training and evaluation of supervised classifiers. At 15 μgL^{-1} , a
436 moderate class imbalance emerged, with 96.5% of samples below the threshold and only 3.5% above.
437 This reflects the challenge of detecting low-prevalence but potentially concerning concentrations,
438 requiring models with high sensitivity for regulatory screening. At 30 μgL^{-1} , the imbalance became
439 extreme: 99.8% of samples were below the threshold, with only 0.2% exceeding it. Such severe
440 imbalance renders conventional binary classification ineffective, as models tend to bias toward the
441 majority class, necessitating specialized techniques for rare-event detection.

442
443 **Performance at the 2 μgL^{-1} threshold**

444
445 Table 3 presents the classification results at the 2 μgL^{-1} threshold. Among all models evaluated,
446 CatBoost achieved the best overall performance, with an F1-score of 80.8% and accuracy of 75.0%,
447 demonstrating an effective balance between precision and recall. Other ensemble methods, including
448 XGBoost, LightGBM, and Random Forest, also performed strongly, with accuracies exceeding 70%.
449 In contrast, simpler models such as the Decision Tree and Multilayer Perceptron (MLP) neural
450 network yielded lower performance, falling slightly below the 70% accuracy threshold. These results
451 confirm that advanced ensemble methods are well-suited for early detection of elevated uranium

levels at lower thresholds, where balanced data allow for reliable model training and meaningful performance assessment.

Table 3 Threshold Classification Performance of 2 μgL^{-1}

Model	Accuracy	Precision	Recall	F1-score	AUC
CatBoost	0.750	0.726	0.915	0.808	0.784
LightGBM	0.736	0.707	0.930	0.803	0.773
XGBoost	0.740	0.708	0.937	0.806	0.772
RandomForest	0.714	0.760	0.740	0.750	0.787
Decision Tree	0.695	0.692	0.853	0.764	0.756
Neural Network(MLP)	0.691	0.736	0.727	0.732	0.777

Table 4 Best model (CatBoost) metrics with 95% Confidence Intervals

Metric	Value	95% CI
F1-score	0.808	[0.787,0.828]
PR-AUC	0.801	[0.764,0.830]
MCC	0.488	[0.444,0.534]
Balanced Accuracy	0.720	[0.697,0.743]
Brier Score	0.185	[0.172,0.198]
ROC-AUC	0.784	[0.760,0.808]

Complete performance metrics with 95% confidence intervals for the best model (**CatBoost**) are provided in Table 4. In addition to comparing the F1 scores, we conducted pair-wise McNemar's test for the model selection. The results show that the CatBoost (F1=0.808) outperformed LightGBM ($\chi^2 = 6.51$, $p = 0.011$), XGBoost ($\chi^2 = 10.38$, $p = 0.001$), Decision Tree ($\chi^2 = 16.95$, $p < 0.001$), and MLP ($\chi^2 = 18.67$, $p < 0.001$) which agrees with the model selection based on F1-score ranking and class imbalance handling characteristics.

Performance at the 15 μgL^{-1} threshold

Table 5 presents model performance at the 15 μgL^{-1} threshold, which reflects a moderate regulatory guideline and introduces a class imbalance of 96.5% (negative) vs. 3.5% (positive).

Table 5 Threshold Classification Performance of 15 μgL^{-1}

Model	Accuracy	Precision	Recall	F1-score	AUC
LightGBM	0.978	0.683	0.689	0.666	0.919
CatBoost	0.975	0.740	0.600	0.632	0.913
XGBoost	0.973	0.689	0.622	0.613	0.907
Random Forest	0.961	0.444	0.444	0.444	0.819
Decision Tree	0.915	0.190	0.444	0.267	0.823
Neural Network(MLP)	0.932	0.522	0.580	0.549	0.854

While most models achieved high overall accuracy (>90%), their F1-scores were moderate, constrained by the scarcity of positive cases and the inherent precision-recall trade-off in imbalanced classification. Among all models, LightGBM achieved the highest F1-score of 66.6%, demonstrating superior ability to balance precision and recall under moderate imbalance. We further performed McNemar's test comparing other models against the LightGBM. The LightGBM (F1 = 0.666) showed no significant difference with CatBoost ($p = 0.458$), but significantly outperformed XGBoost ($\chi^2 = 6.04$, $p = 0.014$), Random Forest ($\chi^2 = 21.13$, $p < 0.001$), Decision Tree ($\chi^2 = 7.90$, $p = 0.005$), and MLP ($\chi^2 = 713.95$, $p < 0.001$). The LightGBM is selected based on slightly higher F1-score for the current case. This performance highlights LightGBM's effectiveness in capturing subtle patterns

associated with elevated uranium levels, even when minority class representation is limited. However, the results also reveal a persistent challenge: even high-accuracy models struggle to consistently detect rare positive instances, underscoring the need for careful threshold tuning and performance metric selection (e.g., F1-score, PR-AUC) over accuracy alone in such scenarios.

Complete performance metrics with 95% confidence intervals for the best model (LightGBM) are provided in Table 6.

Table 6 Best model (LightGBM) metrics with 95% Confidence Intervals

Metric	Value	95% CI
F1-score	0.678	[0.563,0.780]
PR-AUC	0.556	[0.402,0.691]
MCC	0.667	[0.550,0.771]
Balanced Accuracy	0.838	[0.769,0.902]
Brier Score	0.029	[0.022,0.036]
ROC-AUC	0.922	[0.880,0.960]

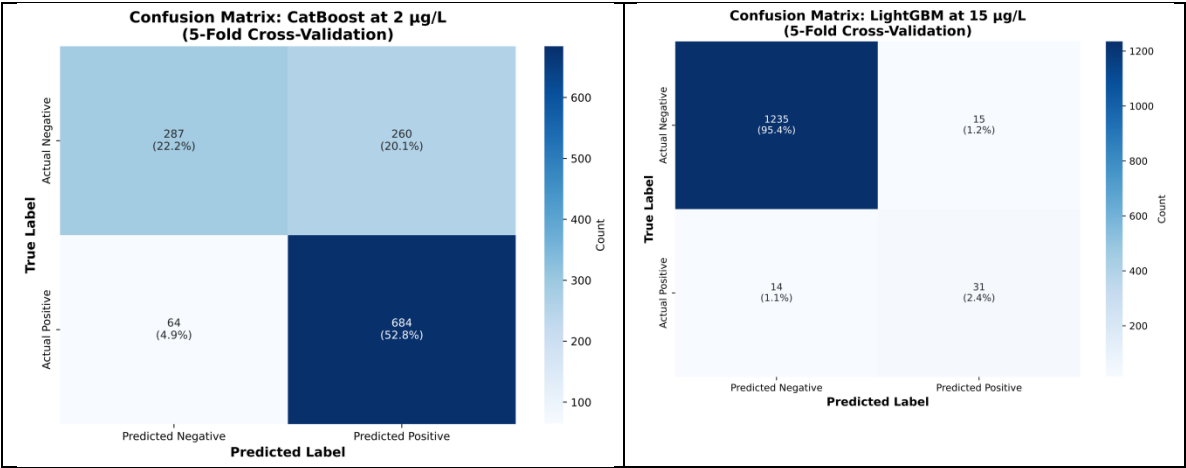


Fig. 5. Confusion Matrices

Confusion matrices for the best-performing models presented in Fig 5. These matrices aggregate predictions across 5-fold cross-validation, providing a comprehensive view of classification performance.

Performance at 30µgL⁻¹ threshold: Rare Event Detection via One-Class Learning

At the 30µgL⁻¹threshold, only 2 out of 1,295 samples (0.154%) exceeded the limit, resulting in extreme class imbalance. Under these conditions, conventional binary classifiers failed to detect any positive cases, achieving near perfect accuracy (>99%) but zero recall and F1-score for the minority class due to complete misclassification of rare events. To address this challenge, we reframed exceedance detection as a high-recall screening task using one-class anomaly detection methods: Isolation Forest and One-Class SVM. These models were trained exclusively on "normal" (non-exceedance) samples and tasked with identifying deviations as potential anomalies. Both models achieved perfect recall (1.00) on the test set, successfully flagging all true exceedance cases with zero false negatives a critical requirement for environmental safety screening. However, their efficiency in prioritizing true positives differed significantly. 30µgL⁻¹ classification performance is presented in Table 7.

- Isolation Forest demonstrated superior performance, providing a 78.3× lift over random screening, meaning it is nearly 78 times more effective at identifying true positives.

- The Number Needed to Review (NNR) metric further illustrates operational efficiency: only ~8 flagged samples need to be verified to identify one true positive using Isolation Forest, compared to ~23 for One-Class SVM.

Although the F1-score for Isolation Forest was low (0.168), this reflects the extreme rarity of positives, and the impact of false positives on precision is not a failure of the model. In this context, perfect recall and high lift are more operationally relevant metrics than F1-score. This approach successfully eliminated missed violations, capturing 100% of the true threshold exceedances a critical requirement for environmental safety screening. While the extreme rarity of positives naturally leads to some false alarms, the model drastically minimizes the overall laboratory workload by maintaining an exceptionally low false positive rate of 1.12%. This translates to a 98.7% reduction in the required screening pool, dropping the number of samples needing physical analysis from the full batch of 1,295 down to an average of just 16.5. Furthermore, the model proved 78.3 times more effective than random sampling.

Table 7 Threshold Classification Performance of 30µg/L⁻¹

Model	F1-score	Precision	Recall	Lift (x)	NNR(1/PREC)	Work load reduction
Isolation forest	0.168	0.121	1.00	78.35	8.26	98.7%
One-class SVM	0.083	0.043	1.00	27.84	23.26	96.4%

Hyper parameter settings obtained for best performance models are given below.

At 2µg/L⁻¹ CatBoost achieved the best performance with optimal hyperparameters: depth=8, learning rate=0.0284, l2_leaf_reg=5.28.

At 15µg/L⁻¹ LightGBM performed best with optimized hyper parameters: num_leaves=22max depth=5, min_data_in_leaf=103, learning_rate=0.190, feature_fraction=0.813.

At 30 µg/L⁻¹ the optimal model shifted to Isolation Forest with tuned hyperparameters: n_estimators=51, max_samples=0.979, contamination=0.0245, max_features=0.834.

Further Analysis and Validation Metrics for Binary Classifiers

To evaluate the discriminative performance of the best-performing models across different operational contexts, we analyzed Receiver Operating Characteristic (ROC) curves at both the 2µg/L⁻¹, 15µg/L⁻¹ thresholds (Fig. 6). The left panel displays the performance of CatBoost at the 2µg/L⁻¹ threshold the optimal model for this relatively balanced classification task. It achieves an AUC of 0.743, indicating good discriminatory power in distinguishing between low and elevated uranium concentrations. While not exceptionally high, this AUC is meaningful in a real-world setting where natural geochemical variability introduces uncertainty. The right panel presents LightGBM performance at the 15µg/L⁻¹ threshold, where class imbalance is more pronounced (~3.5% positive cases). Despite this challenge, LightGBM achieves a strong AUC of 0.910, demonstrating excellent ability to rank samples with true exceedances higher than non-exceedances. This high AUC underscores the model's robustness in detecting rare events, even when positive instances are scarce. ROC curves are particularly valuable in imbalanced classification scenarios because they assess model performance across a continuous range of decision thresholds, rather than at a single fixed point. This enables flexible deployment: decision makers can select operating points that align with specific policy or operational needs, such as prioritizing high recall to ensure no high-concentration samples are missed (critical for public health), or controlling false positive rates to manage laboratory testing capacity and costs. The Area under the ROC Curve (AU-ROC) provides a threshold-independent measure of a model's ranking capability, making it more reliable than accuracy or F1-score in imbalanced settings. A high AU ROC indicates that the model consistently assigns higher probabilities to positive cases, regardless of the final classification threshold. These results confirm that both models, particularly LightGBM at the 15µg/L⁻¹

threshold, possess strong discriminative capacity, supporting their use in real-world uranium screening programs where operational flexibility and reliable risk stratification are essential.

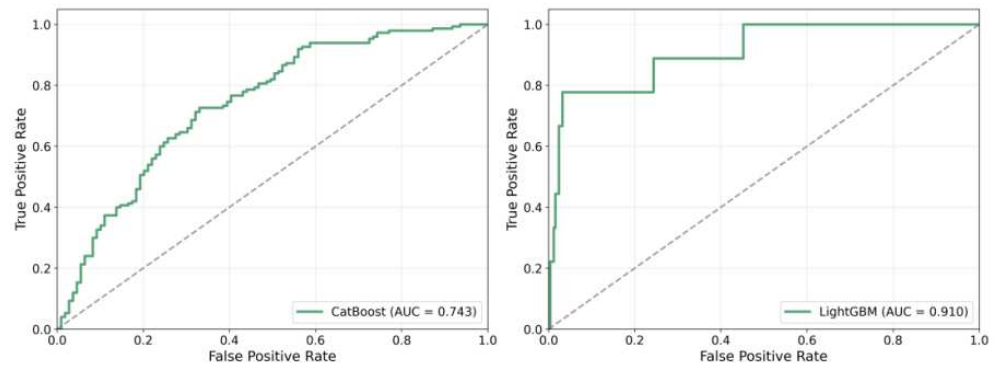


Fig. 6. Receiver Operating Characteristic (ROC) curves for binary classification

The calibration plots demonstrate how well the predicted probabilities align with actual outcome frequencies Fig 7. The PR curves with 95% confidence intervals provide a more informative assessment than ROC curves for imbalanced classification tasks Fig 8. The calibration curve for $2\mu\text{g/L}^{-1}$ (CatBoost) model closely follows the diagonal perfect calibration line, indicating well-calibrated probability estimates. The uncertainty bands (derived from 1000 bootstrap iterations) are relatively narrow, reflecting stable calibration across different data subsamples. The PR curve demonstrates strong precision-recall trade-off across all operating points, with narrow confidence bands indicating robust performance.

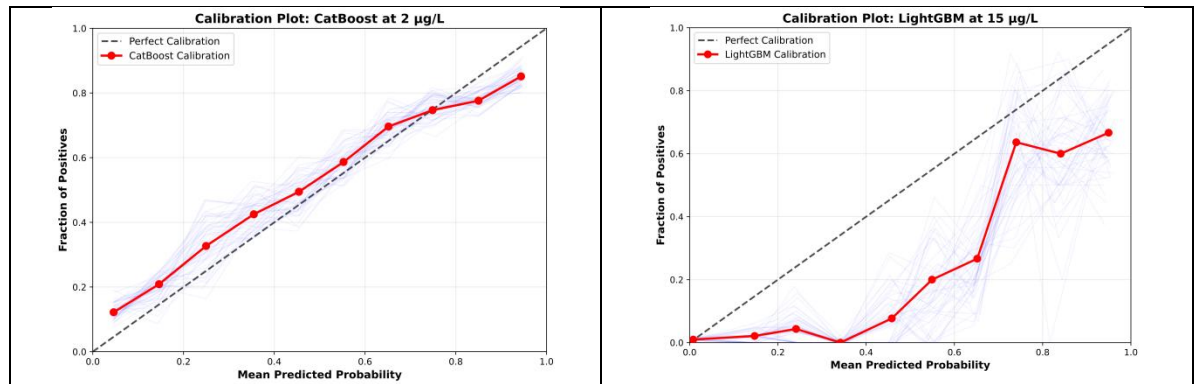


Fig. 7. Calibration Plots

For the $15\mu\text{g/L}^{-1}$ (LightGBM) model, despite a significant class imbalance (96.5% negative), the model shows good calibration in the mid-probability range. The slightly wider uncertainty bands reflect the increased variance expected with fewer positive samples. The wider confidence intervals in the PR curve reflect the challenge of detecting the minority class (3.5% prevalence) for the highly imbalanced dataset. Both models exhibit good calibration, validating the reliability of predicted probabilities for risk stratification. PR curves confirm strong performance despite class imbalance, with PR-AUC significantly exceeding random baseline. Narrow uncertainty bands for $2\mu\text{g/L}^{-1}$ and wider bands for $15\mu\text{g/L}^{-1}$ appropriately reflect the respective data characteristics. These visualizations support the operational deployment of these models for groundwater screening applications.

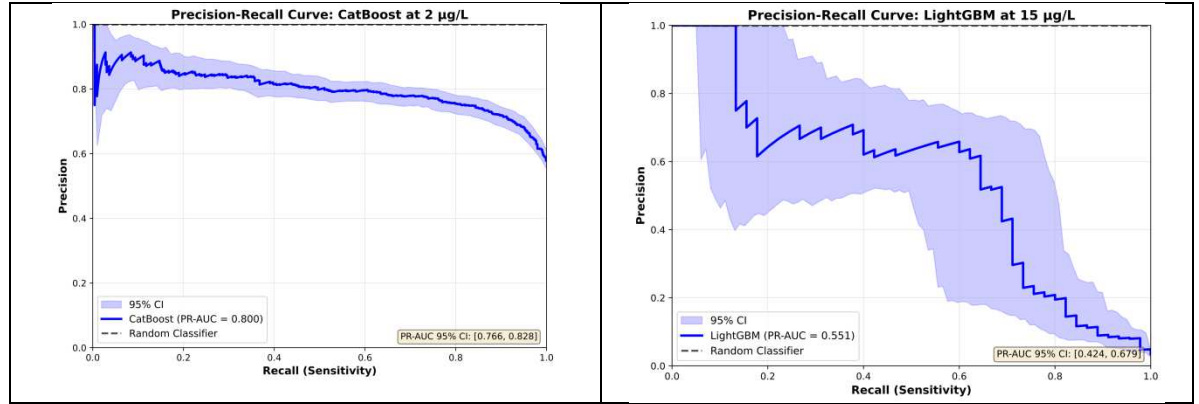


Fig. 8. Precision Recall Curves Plots

Feature Importance Analysis

To identify the physicochemical parameters most influential in predicting uranium exceedance, we employed an ensemble of complementary feature attribution methods: Random Forest feature importance, permutation importance, and SHAP (SHapley Additive exPlanations). Using multiple interpretability techniques enhances robustness and ensures consistency in identifying key predictors, reducing reliance on any single method's assumptions or biases. The top-ranked features across all approaches are summarized in Table 8.

**Table 8 Feature importance rankings from multiple methods
(water quality parameters only)**

Rank	Feature	RF importance	Permutation	SHAP	Mean
1	TDS	15.6%	14.2%	13.8%	14.5%
2	Hardness	14.3%	12.8%	12.5%	13.2%
3	Chloride	13.1%	11.5%	11.2%	11.9%
4	Sulphate	11.8%	9.8%	9.5%	10.4%

Total Dissolved Solids (TDS) emerged as the most important predictor, with a mean importance of 14.5% across methods. This reflects its strong association with water rock interaction intensity and long groundwater residence times, both well established drivers of uranium mobilization from mineral matrices into aqueous phases. Hardness ranked second (13.2%), consistent with its geochemical role in influencing uranium solubility. Hardness, primarily due to calcium and magnesium ions, promotes the formation of stable uranyl carbonate complexes, which enhance uranium mobility under oxidizing, carbonate-rich conditions. Chloride exhibited substantial influence (11.9%), likely due to its contribution to ionic strength and potential for competitive complexation, which can alter uranium speciation and desorption behavior. Sulphate also contributed meaningfully (10.4%), suggesting involvement in ion exchange processes and the dissolution of uranium bearing minerals under variable redox conditions. The strong agreement across independent interpretability methods, not only in ranking but also in relative importance, supports the reliability of these findings. More importantly, the alignment between data driven feature rankings and established hydrogeochemical principles reinforces the scientific plausibility of the model's predictions. This convergence of machine learning interpretability and domain knowledge enhances confidence in the model's internal logic and supports its use as a tool for environmentally informed decision-making, such as prioritizing monitoring efforts or guiding remediation strategies based on measurable water quality indicators.

Spatial Validation

To more rigorously evaluate spatial generalizability, we implemented spatial cross-validation (SCV), in which geographically distinct regions are held out during training and used exclusively for testing. As shown in Fig. 9 (middle panel), the study area was partitioned into five spatially contiguous folds

628 using a clustering-based strategy that preserves geographic coherence. This design better reflects real-
 629 world deployment scenarios, where models trained in one region must reliably predict outcomes in
 630 previously unseen locations. The right panel of Fig. 9 compares F1-scores under standard (random)
 631 and spatial cross-validation across thresholds. At the $2\mu\text{g L}^{-1}$ threshold, the model achieved an F1-score
 632 of 0.803 under standard CV and 0.785 under spatial CV a marginal decrease of approximately 2.2%.
 633 Similarly, at the $15\mu\text{g L}^{-1}$ threshold, performance declined slightly from 0.666 (standard CV) to 0.642
 634 (spatial CV). These small performance drops confirm that the model maintains strong predictive
 635 capability even when extrapolating to spatially independent regions. The results demonstrate robust
 636 geographic transferability, indicating that the model is not merely memorizing local patterns but has
 637 learned generalizable hydrogeochemical relationships. Collectively, these findings validate the
 638 model's practical applicability beyond the training domain. The ability to perform reliably in spatially
 639 unseen areas strengthens confidence in its use for large scale uranium exceedance monitoring and
 640 targeted screening programs, particularly in regions with limited direct uranium measurements.
 641 Comprehensive, spatial visualizations were shown in (Fig. 10a and Fig 10b). Fig 10a Spatial Cross-
 642 Validation Blocks Map, which displays the 5 K-means clusters overlaid on a clean CartoDB Positron
 643 base map. Semi-opaque convex hulls to clearly delineate the geographically distinct partitions used for
 644 spatial validation were added. Fig. 10b. shows the Uranium distribution map, to visualize the actual
 645 uranium concentrations using a continuous color map, providing a clear and accurate representation
 646 of spatial distribution without visual clutter. Five Spatial blocks are created using k-means clustering
 647 for spatial cross-validation metrics which help in finding out spatial generalizability of the trained
 648 model across the region. Spatial cross-validation generally yielded more conservative performance
 649 estimates, with modest reductions (typically $<5\%$) compared to standard cross-validation, indicating
 650 that our models maintain robust predictive capability when extrapolating to new geographic regions.
 651 The spatial validation framework provides confidence that the observed model performance
 652 generalizes beyond the specific sampling locations in our dataset. Uranium Distribution map visualizes
 653 the actual uranium concentrations using a continuous color map, providing a clear and scientifically
 654 accurate representation of spatial distribution without visual clutter.
 655

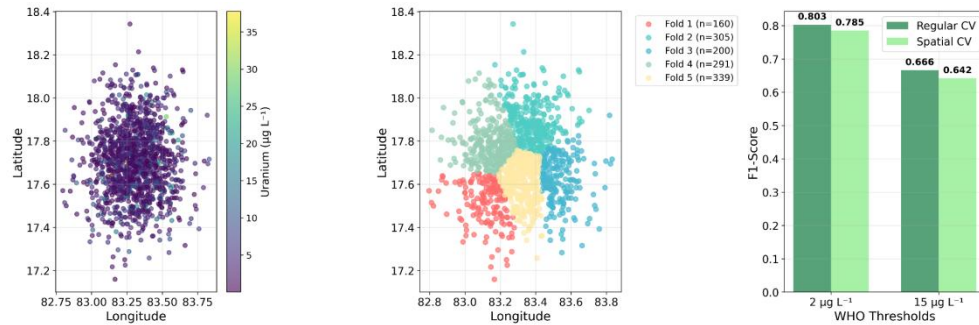


Fig. 9. Spatial distribution and validation

Left: Uranium concentrations across the study area.

Middle: Spatial cross validation folds (n=5) preserving geographic contiguity.

Right: Comparison of F1-scores regular and spatial CV across 2, $15\mu\text{g L}^{-1}$ thresholds.

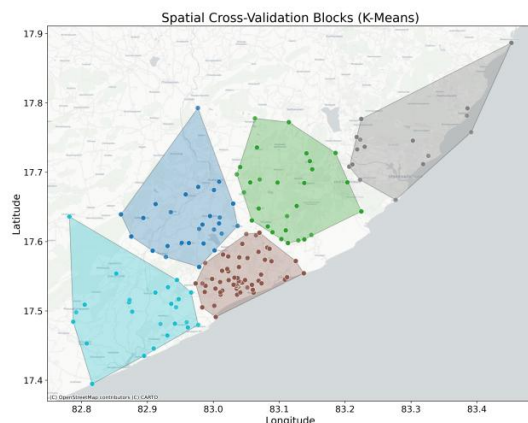


Fig. 10 a. Blocks for spatial validation

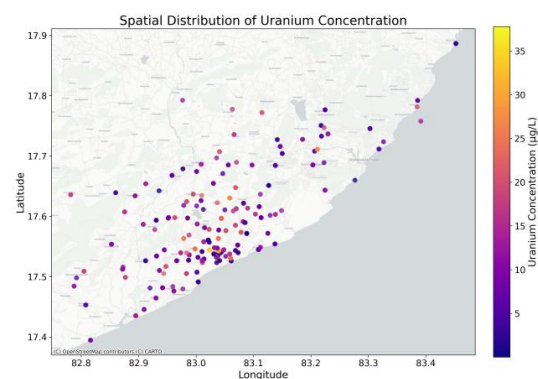


Fig 10 b. Spatial distribution map of Uranium

Discussions

The results demonstrate that physicochemical parameters can effectively predict uranium occurrence in groundwater, reflecting their control over underlying hydrochemical processes. The high predictive performance achieved, particularly at lower concentration thresholds (F1-score: 80.8% at $2 \mu\text{gL}^{-1}$), indicates that routine water quality parameters capture the dominant geochemical signatures governing uranium mobility. Evidence from earlier studies (Nikitenkov A.N. et al. 2025) indicates that uranium mobility can vary across a wide geochemical range, reinforcing the premise of the present study that routinely measured physicochemical parameters can be effectively utilized as predictors for uranium concentration through data-driven modeling approaches. SHAP-based feature attribution identifying pH, TDS, and bicarbonate as dominant predictors is in agreement with uranium geochemical behavior, including pH-dependent solubility, carbonate complexation, and ionic strength effects governing uranium speciation and transport (Langmuir, 1978; Grenthe et al., 1992).

The successful application of one-class anomaly detection (Isolation Forest) at extreme class imbalance ($30\mu\text{gL}^{-1}$ threshold) represents a methodological advancement over traditional binary classifiers, which fail completely under such conditions. The low false positive rate (1.12%) and ~98.7% reduction in screening workload highlight the model's ability to substantially minimize laboratory effort while maintaining complete detection of exceedances. These findings demonstrate that anomaly detection, particularly Isolation Forest, provides a robust and operationally efficient early-warning framework for rare-event environmental monitoring.

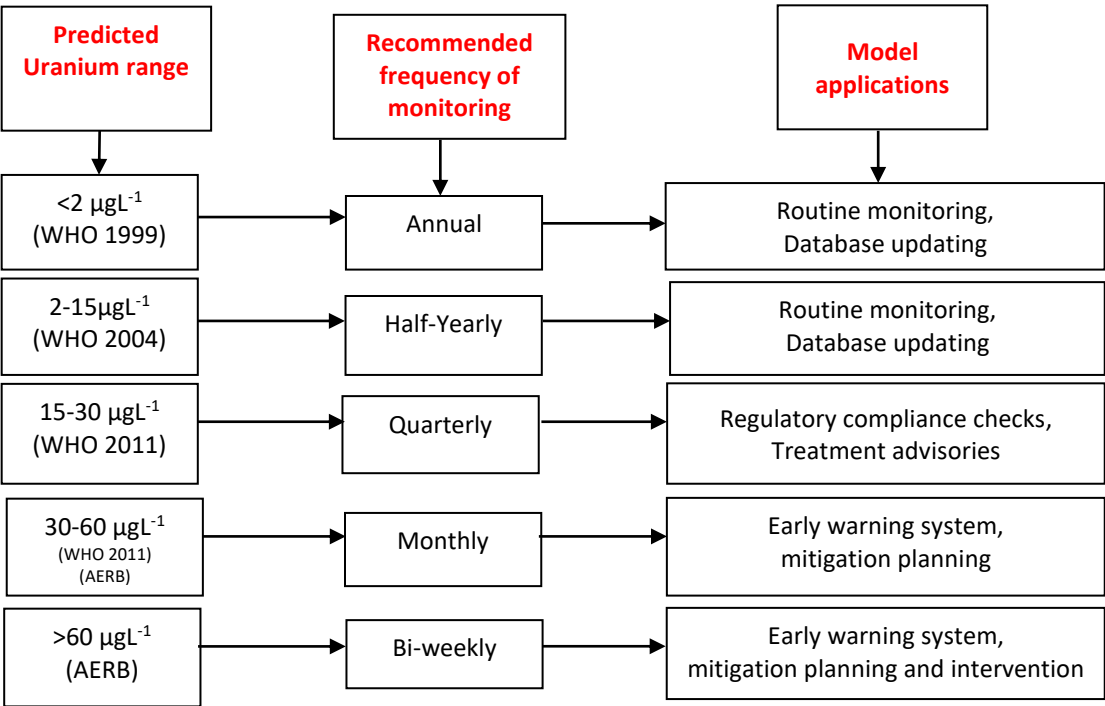
The three uranium concentration thresholds (2, 15, and $30\mu\text{gL}^{-1}$) are based on the evolution of WHO drinking water guidelines (1998, 2004, and 2011, respectively), providing a scientifically grounded framework for hierarchical health assessment. We note that the provided calibration curves enable end-users to adjust decision thresholds based on their specific operational requirements. For instance, monitoring programs prioritizing high sensitivity (e.g., pre-operational baseline surveys) can select lower probability thresholds to maximize recall at the cost of increased false positives, while resource-constrained programs may choose higher thresholds to minimize laboratory confirmation costs. The well-calibrated probability estimates ensure that such threshold adjustments yield predictable changes in precision-recall trade-offs, supporting flexible deployment across diverse operational contexts.

The cited studies (Knoll et al., 2019; Nafouanti et al., 2021; Gupta et al., 2022; Podgorski et al., 2022; Yang et al., 2023; Anjali et al., 2024) primarily applied machine learning to common groundwater contaminants such as fluoride, nitrate, chloride, or arsenic, whereas the present study specifically focuses on uranium, which exhibits distinct redox and salinity-controlled geochemical behavior. Although (Karunanidhi et al., 2025, Durga Prasad Panday et al. 2024) modeled uranium concentrations using regression frameworks evaluated by R^2 and RMSE, our study adopts a threshold-

697 based classification approach aligned with precautionary and regulatory limits (2, 15, and 30 μgL^{-1})
 698 and evaluates performance using F1-score under spatial cross-validation. Additionally, the integration
 699 of SHAP-based interpretability and the use of one-class anomaly detection to address extreme class
 700 imbalance at higher thresholds distinguish the present work from earlier uranium prediction studies,
 701 advancing it toward a risk-informed operational screening framework.
 702

703 Practical Applications

704
 705 This study presents a tiered, risk-informed framework for predicting uranium concentrations and
 706 guiding adaptive groundwater monitoring. Predicted values are categorized into regulatory and
 707 scientific thresholds to enable differential surveillance intensity.
 708



732
 733 **Fig. 11. Derivation of Recommendations & Functional Applications**

734
 735 As shown in Fig. 11, low to moderate concentrations (<2 μgL^{-1} and 2–15 μgL^{-1}) support baseline and
 736 routine monitoring, whereas elevated levels ($\geq 30\mu\text{gL}^{-1}$, World Health Organization guideline; ≥ 60
 737 μgL^{-1} , Atomic Energy Regulatory Board limit) trigger compliance verification, confirmatory testing,
 738 and source investigation. Thresholds were defined using a cost-sensitive, performance-based
 739 framework that integrates detection objectives with operational constraints, including analytical cost,
 740 laboratory capacity, sampling logistics, and regulatory reporting timelines. The design prioritizes high
 741 sensitivity for exceedance detection while controlling false positives to avoid unnecessary testing and
 742 regulatory burden, ensuring scalable and resource-efficient implementation. Applicable across all
 743 facility life-cycle phases, the approach leverages routinely measured physicochemical parameters to
 744 convert sparse uranium analyses into spatially and temporally dense risk assessments, strengthening
 745 monitoring efficiency and regulatory defensibility.

746 Thus this study presents a comprehensive comparison of advanced machine learning
 747 algorithms, including ensemble methods and one-class learners, for uranium concentration
 748 classification in groundwater. The integration of multi-threshold classification, spatial validation,
 749 SHAP interpretability, and practical deployment guidelines into a unified framework represents a

significant advancement in environmental ML applications. Fig. 1 provides a replicable blueprint for implementing ML based groundwater monitoring in other regions facing similar challenges.

Model Limitations and Future Research Directions

Despite promising performance, the proposed ML framework has several limitations. The dataset (~1,295 samples) provides limited spatial coverage and a moderate range of uranium concentrations, which may restrict generalizability, particularly at concentration extremes. Ensemble models also exhibit inherent “black-box” behavior, and interpretability methods may not fully resolve complex feature interactions. In addition, the assumption of stationary environmental conditions may reduce long-term predictive reliability.

Future work should expand datasets and feature sets, explore deep-learning approaches for spatial modeling, develop adaptive real-time prediction systems, and extend the framework to multi-radionuclide or integrated water-quality assessment, thereby improving robustness and operational applicability.

Conclusions

This study integrates physicochemical parameters and data driven modeling to predict uranium occurrence in groundwater, highlighting the role of hydrogeochemical controls. The results indicate that reliable predictive performance can be achieved without explicit spatial or temporal inputs, suggesting that physicochemical parameter’s ability to capture key processes such as mineral dissolution, salinity evolution, and carbonate complexation. The proposed multi-threshold classification framework enables assessment across precautionary, guideline, and regulatory concentration limits, supporting flexible application in groundwater monitoring and risk assessment. Ensemble learning methods effectively represent nonlinear interactions among variables, while the use of one-class anomaly detection addresses challenges associated with severe class imbalance at higher concentration thresholds. The study introduces a structured threshold-based prediction strategy combined with spatially independent validation and interpretable machine learning using SHAP, enhancing model transparency and environmental relevance. These features improve confidence in the model’s applicability for decision-support in groundwater monitoring. Importantly, the consistency between model-derived feature importance and established hydrogeochemical understanding confirms that the predictive framework is both interpretable and scientifically meaningful. From an application perspective, the framework offers potential for scalable deployment, including adaptation to other regions and integration into monitoring programs where rapid screening is required. Its ability to identify elevated uranium concentrations supports proactive management and prioritization of detailed investigations. While the results are promising, further validation with larger and more diverse datasets, including hydrogeological and temporal variables, would strengthen model generalizability. Overall, the proposed approach provides a practical and interpretable tool for supporting groundwater quality assessment and regulatory decision-making in resource-constrained and data-limited settings.

References:

- AERB (2004) Limits on Uranium in Drinking Water: AERB/VC/16/2004/80/373, 29th June 2004.
- Akiba T, Sano S, Yanase T, Oht, T, Koyama M, (2019) Optuna: A next-generation hyper parameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2623-2631.
- Alinne Marianne Martins Araújo, Laercio Leal dos Santos, BeataKozłowska, AgataWalencik-Łata, Ricardo de Aragão, de Paiva William, Antônio Augusto Pereira Sousa, AlanKellnonNóbregaCarvalho, Josenildo Isidro dos Santos Filho, ThyagoNobregaSilveira, Elibe Silva Souza, Susana de Souza Lalic; Integrated hydrochemical and radiological assessment of groundwater in the Seridó crystalline basement (Brazil): Evidence of geogenic influences on Ra and trace elements. Journal of Environmental Radioactivity Vol:293, Feb 2026, 107877, <https://doi.org/10.1016/j.jenvrad.2025.107877>.

804 Anjali Kerketta, Harmanpreet Singh Kapoor, Prafulla Kumar Sahoo, Groundwater fluoride
805 prediction modeling using physicochemical parameters in Punjab, India: a machine-learning
806 approach. <https://doi:10.3389/fsoil.2024.1407502>.

807 Anjali Kerketta, Harmanpreet Singh Kapoor, Prafulla Kumar Sahoo, Machine learning based
808 prediction of elevated uranium concentrations in shallow groundwater of Punjab, India by
809 Environ Geochem Health. 2026 Feb 17;48(4):166. <https://doi:10.1007/s10653-026-03004-2>.

810 Ayotte, J.D., Nolan, B.T., et al. (2011) Modeling the probability of arsenic and uranium
811 concentrations in private wells in the United States. Environmental Science & Technology, 45,
812 3218–3225.

813 Balaram V,Rani A,Rathore D.P.S.,Uranium in groundwater in parts of India and world: A
814 comprehensive review of sources,impact to the environment and human health, analytical
815 techniques and mitigation technologies. Geosystems and Geoenvironment 1(2022) 100043.

816 Belgiu M, Drăguț L, (2016) Random forest in remote sensing: A review of applications and future
817 directions: ISPRS Journal of Photogrammetry and Remote Sensing, 114:24-31

818 Bergstra J, Bengio Y, (2012) Random search for hyper-parameter optimization: Journal of Machine
819 Learning Research, 13(Feb), 281-305.

820 BIS (2012) Indian Standard Drinking Water Specification (Second Revision): Bureau of Indian
821 Standards (IS10500:2012). https://cpcb.nic.in/wqm/BIS_Drinking_Water_Specification.pdf

822 BIS (2021) Amendment No. 3 February 2021 to IS 10500:2012 drinking water- specification (Second
823 Revision). Bureau of Indian Standards: <https://www.scribd.com/document/656107401/10500A3>.

824 Breiman L.(2001) Random forests. Machine Learning, 45(1), 5-32.

825 Chen K, Chen H, Zhou C, Huang Y, Qi X, Shen R, Zhang Y, (2018) Comparative analysis of surface
826 water quality prediction performance and identification of key water parameters using different
827 machine learning models based on big data: Water Research, 171:115454.

828 Chen T, Guestrin C, (2016) XGBoost: A scalable tree boosting system. In Proceedings of the 22nd
829 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: pp 785-
830 794.

831 Chinchor N, (1992) MUC-4 evaluation metrics: In Proceedings of the 4th Conference on Message
832 Understanding, pp 22-29.

833 Cortes C, Vapnik V, (1995) Support-vector networks: Machine Learning, 20(3), 273-297.

834 Coyte, R.M., Singh, A., Furst, K.E., Mitch, W.A., Vengosh, A., 2018. Large-scale uranium
835 contamination of groundwater resources in India. Environmental Science & Technology Letters
836 5(6), 341-347. <https://doi.org/10.1021/acs.estlett.8b00215>

837 Dougherty J, Kohavi R, Sahami M, (1995) Supervised and unsupervised discretization of continuous
838 features: In Machine Learning Proceedings 1995 pp 194-202.

839 Durga Prasad Panday, Manish Kumar, (2024) Decoding uranium variation over the Indian Peninsula
840 through leveraging standardized precipitation evapotranspiration indices and groundwater level
841 fluctuations Groundwater for Sustainable Development: Volume 26, August 2024, 101241
842 Research paper. <https://doi.org/10.1016/j.gsd.2024.101241>.

843 Efron B, (1979) Bootstrap methods: another look at the jackknife. The Annals of Statistics, 7(1), 1-26.

844 Grenthe, I., Fuger, J., Konings, R.J.M., Lemire, R.J., Muller, A.B., Nguyen-Trung, C., Wanner, H.,
845 1992.Chemical Thermodynamics of Uranium.North-Holland, Amsterdam.

846 Gupta P K, Maiti S, (2022) Enhancing data-driven modeling of fluoride concentration using new data
847 mining algorithms: Environ Earth Sci. 81:89.<https://doi:10.1007/s12665-022-10216-z>.

848 Guyon, I.,&Elisseff,A.(2003). An introduction to variable and feature selection. Journal of Machine
849 Learning Research, 3, 1157–1182.

850 Käßner, H. T. Kalapurakkal, B.Huber, M. Tichomirowa; A New Water-Based 87Sr/86Sr Isoscape
851 Map of Central and NEGermany, with Special Emphasis on mountainous Regions; Aquatic
852 Geochemistry (2023) 29:95–125 <https://doi.org/10.1007/s10498-023-09412-5>.

853 Karunanidhi D,Raj,M.R.H.,Roy,P.D. Integrated machine learning based groundwater quality
854 prediction through groundwater quality index for drinking purposes in a semi arid river basin of
855 South India .Environmental Geochemistry and Health 47,119(2025)
856 <https://doi.org/10.1007/s10653-025-02425-9>

857 Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu T Y, (2017) LightGBM: A highly efficient
858 gradient boosting decision tree: Advances in Neural Information Processing Systems, 30:3146-
859 3154.

860 Knoll L, Breuer L, Bach M, (2019) Large scale prediction of groundwater nitrate concentrations from
861 spatial data using machine learning: Science Total Environ 668:1317–27. [https://doi:
862 10.1016/j.scitotenv.2019.03.045.](https://doi.org/10.1016/j.scitotenv.2019.03.045)

863 Kohavi R, (1995) A study of cross-validation and boot strap for accuracy estimation and model
864 selection. In IJcai Vol.14 No.2 pp1137-1145.

865 Kumar S. Yadav, A.L.Ramanathan, Manojkumar, Chidambaram S. Gautam Y.P., Tiwar C. Assessment
866 Of arsenic and uranium co-occurrences in ground ater of central
867 Gangetic Plain, Uttarpradesh, India. Environmental Earth Sciences (2020) 79:154.
868 <https://doi.org/10.1007/s12665-020-8892-x>

869 Langmuir, D., 1978. Uranium solution-mineral equilibria at low temperatures with applications to
870 Sedimentary ore deposits. Geochimica et Cosmochimica Acta 42(6), 547-
871 569. [https://doi:10.1016/0016-7037\(78\)90001-7](https://doi:10.1016/0016-7037(78)90001-7)

872 Nafouanti M B, Li J, Mustapha N A, Uwamungu P, AL-Alimi D, (2021) Prediction on the fluoride
873 contamination in groundwater at the Datong Basin, Northern China: Comparison of random
874 forest, logistic regression and artificial neural network. Appl Geochem 132:105054. [https://doi:
875 10.1016/j.apgeochem.2021.105054.](https://doi:10.1016/j.apgeochem.2021.105054)

876 Nikitenkov A.N, D. A. Novikov, T. V. Korneeva, F. F. Dultsev; Numerical Physicochemical
877 Modeling of the Geochemical Behavior of Uranium and Thorium in Radon Waters of Southern
878 Siberia, Aquatic Geochemistry (2025) 31:6 [https://doi.org/10.1007/s10498-025-09440-3.](https://doi.org/10.1007/s10498-025-09440-3)

879 Nolan, J., Weber, K.A., 2015. Natural uranium contamination in major U.S. aquifers linked to nitrate.
880 Environmental Science & Technology Letters 2(8), 215-220.
881 [https://doi.org/10.1021/acs.estlett.5b00174.](https://doi.org/10.1021/acs.estlett.5b00174)

882 Padma Savitri P., B.Ramesh, J.Sudhakar, R.Balaram Kumar, U.Bharadwaj,
883 M.Warrier, A.D.P.Rao.

884 A.Vinod Kumar., Radiological and Physicochemical parameters in the aquatic
885 environment at
886 the northeastern coast of Andhra Pradesh, India - A comprehensive study MAPAN-Journal
887 of Metrology Society of India (2026). <https://doi.org/10.1007/s12647-026-00929-y>

888 Podgorski J, Berg M, (2022) Global analysis and prediction of fluoride in groundwater: Nat Commun.
889 13:4232. [https://doi: 10.1038/s41467-022-31940-x.](https://doi:10.1038/s41467-022-31940-x)

890 Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V, Gulin A, (2018) CatBoost: unbiased boosting
891 with categorical features: In Advances in Neural Information Processing Systems pp 6638-6648.

892 Roberts D R, Bahn V, Ciuti S, Boyce M S, Elith J, Guillera-Arroita G, Dormann C F, (2017)
893 Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure:
894 Ecology 40(8) 913-929.

895 Sahoo S K, Jha S K, Jha V N, Patra A C, Kulkarni M S, (2021) Survey of uranium in drinking water
896 sources in India: Interim observations: Current Science, Vol.20, No.9.

897 Saito T, Rehmsmeier M, (2015) The Precision–Recall Plot Is More Informative than the ROC Plot
898 When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432.

899 UNSCEAR (2016a) Sources, effects and risks of ionizing radiation. United Nations Scientific
900 Committee on the Effect of Atomic Radiation, Annex D – Biological effects of selected internal
901 emitters – Uranium, United Nations, New York.

902 Van der Maaten L, Hinton G, (2008) Visualizing Data using t-SNE: Journal of Machine Learning
903 Research, 9, 2579–2605.

904 WHO (1998) Guidelines for drinking Water quality: 2nd Edn, Addendum to volume 2, Health criteria
905 and other supporting information, World Health Organization, Geneva.
906 <https://www.who.int/publications/i/item/WHO-EOS-98.1>

907 WHO (2004) Guidelines for Drinking Water Quality, 3rd Edn. World Health Organization, Geneva.

908 WHO (2012) Uranium in drinking water, Switzerland, Geneva. WHO/SDE/WSH/03.04/118/Rev/1

909 WHO (2022) Guidelines for Drinking Water Quality, 4thEdn. World Health Organization, Geneva.
 910 Yang H, Wang P, Chen A, Ye Y, Chen Q, Cui R, (2023) Prediction of phosphorus concentrations in
 911 shallow groundwater in intensive agricultural regions based on machine learning: Chemosphere
 912 313:137623. <https://doi: 10.1016/j.chemosphere.2022.137623>.

913

914

915

916 **Appendix: Technical Glossary for Environmental Scientists**

917

918 **Machine Learning Terms:**

- 919 • XGBoost (Extreme Gradient Boosting): An advanced ensemble algorithm that builds
 920 multiple decision trees sequentially, with each tree learning from the errors of previous trees.
- 921 • Random Forest: An ensemble method that creates multiple decision trees using random
 922 subsets of data and features then averages their predictions.
- 923 • SHAP (Shapley Additive explanations): A method to explain individual predictions by
 924 calculating each feature's contribution.
- 925 • Cross validation: A technique to assess model generalization by training and testing on
 926 different data subsets.
- 927 • F1-Score: Harmonic mean of precision and recall, providing balanced performance
 928 assessment.
- 929 • AUC-ROC: Area Under the Receiver Operating Characteristic curve, measuring
 930 classification performance across all thresholds.

931 **Statistical Terms:**

- 932 • Coefficient of Variation (CV): Ratio of standard deviation to mean, indicating relative
 933 variability
- 934 • Skewness: Measure of distribution asymmetry (positive = right tail, negative = left tail).
- 935 • Kurtosis: Measure of distribution tail heaviness (high = more outliers).
- 936 • Bootstrap Confidence Intervals: Uncertainty estimates obtained by resampling data multiple
 937 times.
- 938 • Permutation Importance: Feature importance calculated by randomly shuffling each feature
 939 and measuring the performance decrease.

940 **Hydrogeochemical Terms:**

- 941 • Total Dissolved Solids (TDS): Sum of all dissolved minerals in water, indicating overall
 942 water quality.
- 943 • Hardness: Primarily calcium and magnesium content, which affects water chemistry and
 944 uranium complexation.
- 945 • Uranyl Complexes: Dissolved uranium species, particularly $\text{UO}_2(\text{CO}_3)_2^{2-}$ and $\text{UO}_2(\text{CO}_3)_3^{4-}$.
- 946 • Water rock Interaction: Geochemical processes where groundwater dissolves minerals from
 947 aquifer materials.
- 948 • Redox Conditions: Oxidation reduction state affecting uranium solubility (oxidizing
 949 conditions increase mobility).