

Machine Learning-Based Prediction Of Groundwater Fluoride Contamination And Dental Fluorosis Risk In Rural Areas

Gaurav Saxena^{1*}, Dr. Priyank Singhal²

^{1*}Research Scholar, Department of Computer Science, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India

²Associate professor, Department of Computer Science, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India

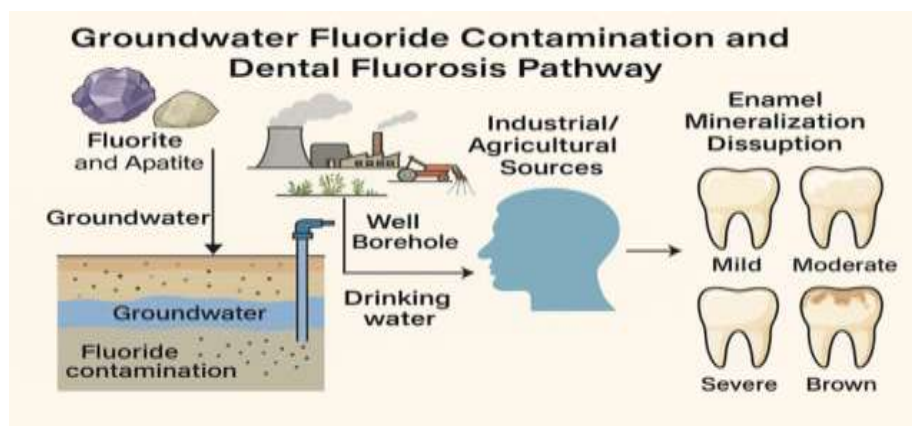
Abstract

Groundwater fluoride contamination remains a major public health concern in rural regions where communities depend on untreated aquifers for drinking water, hindering progress toward Sustainable Development Goals 3 and 6. This study examines groundwater fluoride dynamics and their association with dental fluorosis across ten rural blocks of Unnao district, Uttar Pradesh, India. A five-year dataset (2021–2025) containing physicochemical water-quality parameters and dental health indicators was analysed, and future conditions were forecast for 2026–2030. Supervised machine-learning models, including regression and multi-class classification with ensemble techniques, were applied to estimate fluoride concentration (mg/L), predict fluorosis prevalence (%), and classify disease severity. Data preprocessing, temporal and stratified data splitting, and feature-selection procedures ensured robust model development. The HistGradientBoosting regressor achieved the best regression performance (RMSE = 0.089; R^2 = 0.985), while the XGBoost classifier produced the highest classification accuracy (0.986) and F1-score (0.985). Forecasts indicate increasing fluoride levels and fluorosis risk, highlighting the need for targeted mitigation and improved rural water-management strategies.

Keywords: Fluoride Contamination, Dental Fluorosis, Machine Learning, Water Quality Prediction, Rural Drinking Water, SDG-3 & 6.

1. INTRODUCTION

Groundwater contamination has become an increasingly critical public-health and environmental concern, particularly in rural regions where communities depend almost entirely on untreated aquifers for drinking water. This situation directly conflicts with the United Nations Sustainable Development Goals, especially SDG 3 (Good Health and Well-Being), which calls for reducing illness from environmental hazards, and SDG 6 (Clean Water and Sanitation), which seeks universal access to safe drinking water. Although groundwater is often regarded as a naturally protected resource, international assessments indicate that it remains highly vulnerable to geogenic pollutants. Among these, fluoride is especially significant owing to its irreversible effects on human health. Concentrations exceeding the World Health Organization recommended limit of 1.5 mg L⁻¹ are known to cause dental and skeletal fluorosis and may contribute to broader metabolic and physiological disturbances (WHO, 2017). These international thresholds, alongside national standards adopted in India, the United States, Canada, China and the European Union, are summarised in Table 1, which outlines permitted fluoride concentrations and the scientific rationale underpinning each regulatory framework. Globally, fluoride contamination affects millions of individuals living in regions such as the East African Rift Valley, northern China, Mexico, Pakistan and parts of the Middle East, where geological formations facilitate the mobilisation of fluoride minerals into groundwater (Rango et al., 2015).

Figure 1. Groundwater Fluoride Contamination and Dental Fluorosis Pathway

The WHO/UNICEF Joint Monitoring Programme has repeatedly identified these areas as priority zones requiring urgent safe-water interventions (UNICEF/WHO JMP, 2021). The pathway through which fluoride enters groundwater is conceptualised in Figure 1, illustrating how minerals such as fluorite and apatite dissolve under alkaline conditions, enhanced by high bicarbonate concentrations, prolonged water-rock interaction and anthropogenic influences from agricultural fertilisers and industrial effluents (Reddy et al., 2020). When such contaminated groundwater is abstracted through handpumps and borewells, communities relying on untreated sources face prolonged fluoride intake, ultimately resulting in enamel hypomineralisation and progressive dental fluorosis. India remains one of the most severely affected countries. Reports issued by the Central Ground Water Board (CGWB) reveal that more than 230 districts exceed the national fluoride limit, with widespread implications for rural public health (CGWB, 2020). Assessments under the National Programme for Prevention and Control of Fluorosis (NPPCF) similarly emphasise the prevalence of fluorosis among children and socio-economically vulnerable groups (MoHFW, 2019). Although the Bureau of Indian Standards (IS 10500:2012/2023) specifies a desirable limit of 1.0 mg L^{-1} and an upper permissible limit of 1.5 mg L^{-1} , exceedances persist in several states (BIS, 2012), reinforcing chronic exposure risks closely aligned with international regulatory observations. Dental-fluorosis changes observed in the field, including mottling and discolouration, correspond directly with the Dean's Index categories listed in Table 2 (Dean, 1942). Unnao district in Uttar Pradesh was selected for this investigation based on hydrogeochemical, socio-environmental and administrative considerations supported by official records and field observations. Monitoring data from the Uttar Pradesh Jal Nigam and the State-Level Water Analysis Laboratory, Lucknow, consistently report fluoride concentrations exceeding Bureau of Indian Standards limits in several blocks, while assessments by the Central Ground Water Board classify parts of the district as chemically stressed due to aquifer depletion and declining groundwater quality (CGWB, 2021; UP Jal Nigam, 2022).

Table 1. International Standards and Regulatory Limits for Fluoride in Drinking Water

(Sources compiled from WHO, BIS (2023), EPA, NHMRC, EU, Health Canada, and related national regulations.)

Organization / Standard	Country / Region	Fluoride Limit (mg L^{-1})	Description / Regulatory Basis
-------------------------	------------------	---------------------------------------	--------------------------------

World Health Organization (WHO)	Global	1.5	Guideline value based on lifelong exposure assessments designed to prevent dental and skeletal fluorosis.
Bureau of Indian Standards (BIS IS 10500: 2012 / 2023)	India	Acceptable: 1.0 Permissible: 1.5	Defines national limits considering tropical climate; higher limit allowed in areas without alternative water sources.
Health Canada	Canada	1.5	Established as the Maximum Acceptable Concentration (MAC) to control chronic fluoride intake risks.
U.S. Environmental Protection Agency (EPA)	United States	Primary MCL = 4.0; Secondary MCL = 2.0	Primary limit controls skeletal fluorosis; secondary standard manages dental mottling and aesthetic quality.
National Health and Medical Research Council (NHMRC) / Australian Drinking-Water Guidelines	Australia	≤ 1.5 (Target: 0.6–1.1)	Defines optimal fluoridation range to balance caries prevention with exposure safety in varied climates.
European Union (EU Drinking-Water Directive 2020/2184)	EU Member States	1.5	Mandatory parametric value uniformly applied across EU nations to ensure safe drinking-water quality.
World Bank / UNICEF Water Quality Programs	Developing Countries	1.5	Adopts WHO guideline for evaluating fluoride exposure in community-level water-supply projects.
National Health Commission of China (GB 5749-2022)	China	1.0	Reduced national limit reflecting higher water intake and temperature-dependent fluoride exposure in warm climates.

The district's predominantly rural character further increases its vulnerability. According to the Jal Jeevan Mission baseline survey, households across blocks including Asoha, Purwa, Sikanderpur Karan, Sikanderpur Sarausi, Bichhiya, Sumerpur, Hasanganj, Hilauli, Bighapur and Nawabganj rely mainly on handpumps and shallow borewells for drinking water (Government of India, 2019). This reliance on untreated groundwater elevates the risk of prolonged fluoride exposure, a finding corroborated by field observations of enamel discolouration and hypomineralisation in multiple villages

(NPPCF, 2020). To address these concerns, the present study analyses multi-year groundwater quality and dental-health data (2021–2025), comprising key physicochemical parameters and fluorosis indicators, and applies supervised predictive modelling techniques to estimate fluoride concentration, disease prevalence and severity, and to generate forward-looking risk projections. The findings are intended to support evidence-based groundwater management and targeted public health interventions in fluoride-affected rural settings.

Table 2. Dean's Index for the Assessment of Dental Fluorosis

(Adapted and refined to align with scientific reporting standards; values represent generalised enamel response categories.)

Score	Category	Clinical Description	Approx. % of Teeth Affected
0	Normal	The enamel surface appears smooth, glossy, and uniformly cream-white, showing no visible evidence of fluorosis.	0%
0.5	Questionable	Minor deviation in translucency; fine white flecks or pale spots occasionally visible.	<10%
1.0	Very Mild	Small opaque patches are present irregularly, generally covering less than one-quarter of the surface area.	<25%
2.0	Mild	White opaque regions become more pronounced and extensive but still affect less than half of the enamel surface; brown staining is absent.	25–50%
3.0	Moderate	Surface wear or brown discoloration apparent; more than half of the enamel surface displays mottling or loss of lustre.	>50%
4.0	Severe	All enamel surfaces affected with dark brown staining, pitting, and evident hypomineralisation producing an irregular or corroded appearance.	≈100%

1.1 Objectives of the Study

In view of the documented hydrogeochemical vulnerability of Unnao district, the observed burden of dental fluorosis, and the availability of a robust five-year dataset, this study was guided by a set of clearly defined objectives. The first objective was to develop a data-driven predictive framework to model groundwater fluoride dynamics

across rural blocks using advanced machine-learning approaches. The second objective was to forecast groundwater fluoride concentrations, dental fluorosis prevalence and severity for the period 2026–2030, thereby enabling early identification of habitations at elevated risk. The third objective was to identify the most affected gram panchayats, villages and habitations within each block by integrating regression-based estimates of fluoride levels and fluorosis prevalence with classification-based severity predictions for 2026–2030. Finally, the study aimed to provide a scientifically grounded decision-support mechanism for local authorities by integrating predicted fluoride levels, fluorosis outcomes and spatial vulnerability patterns into an operational early-warning framework tailored to fluoride-affected rural communities.

2. Related Work

Groundwater fluoride contamination has been extensively documented as a persistent environmental and public-health challenge across diverse geological regions. Early hydro-geochemical investigations attributed fluoride enrichment primarily to the dissolution of minerals such as fluorite, apatite, biotite and amphiboles under conditions of alkaline pH, high bicarbonate content and prolonged water–rock interaction. Studies undertaken in South Korea, India and China demonstrated that increasing alkalinity, reduced calcium activity and longer groundwater residence times intensify fluoride mobilisation (Chae et al., 2007; Saxena et al., 2018). Subsequent research conducted between 2015 and 2025 consolidated these observations, indicating that climatic variability, aquifer lithology, evapotranspiration intensity and anthropogenic abstraction significantly influence fluoride concentrations. For example, investigations in the East African Rift Valley, northern China, the Indo-Gangetic Plains and Mexico highlighted how fluctuating climatic regimes, irrigation return flows and aquifer heterogeneity shape fluoride distribution patterns (Rango et al., 2016; Liu et al., 2017; Martínez et al., 2021; Zhang et al., 2022). These findings underscore the pronounced spatial heterogeneity of fluoride occurrence and the need for location-specific predictive approaches. India, which hosts some of the world's largest fluoride-affected belts, has contributed a substantial body of empirical research during this period.

The Central Ground Water Board (CGWB) and various state agencies have identified high-fluoride zones in Rajasthan, Telangana, Andhra Pradesh, Karnataka, Madhya Pradesh, Bihar and Uttar Pradesh. Academic studies complement these assessments: Reddy et al. (2019) reported an R^2 of 0.91 for fluoride prediction in Telangana; Kumar et al. (2021) achieved an RMSE of 0.26 mg/L in Bundelkhand aquifers using multivariate models; Ghosh et al. (2024) demonstrated that ensemble models reduced MAE to approximately 0.10 mg/L in parts of Madhya Pradesh; and Dubey et al. (2025) recorded an R^2 of 0.95 with MAE under 0.09 mg/L using CatBoost across multiple Indian states. These results consistently indicate that non-linear, multi-parameter models capture groundwater–fluoride relationships more effectively than conventional regression-based techniques. Parallel research has focused on the epidemiological impact of chronic fluoride exposure. Dental fluorosis—characterised by enamel mottling, discolouration and hypomineralisation—remains widely prevalent in fluorotic regions of India, China, Kenya, Ethiopia and Pakistan. Studies conducted in India, including those by Mandal et al. (2021), Gupta et al. (2022), and Rathore et al. (2023), reported fluorosis prevalence ranging from 28–52% among school-aged children, with severe grades identified in several rural districts. National Programmes for Prevention and Control of Fluorosis (NPPCF) and related health-sector records corroborate these findings, reinforcing the need for forward-looking predictive systems rather than solely retrospective assessments. The past decade has witnessed rapid growth in the application of machine-learning (ML) techniques to groundwater-quality prediction. More than twenty

algorithms have been employed across published studies, ranging from classical linear approaches to advanced ensemble methods. Early works utilised Linear Regression, k-Nearest Neighbours and Support Vector Regression; however, tree-based ensemble learners—such as Random Forest, Extra Trees, Gradient Boosting, XGBoost, CatBoost and LightGBM—have consistently achieved superior performance. Reported values from studies undertaken between 2015 and 2025 demonstrate that these algorithms typically yield $R^2 > 0.90$, $RMSE < 0.25$ mg/L and $MAE < 0.12$ mg/L, markedly outperforming linear and kernel-based models, which often remain confined to R^2 ranges of 0.75–0.85. Histogram-based Gradient Boosting (HistGradientBoosting), an efficient variant of gradient boosting employing binning optimisation, has emerged more recently and demonstrated strong predictive capability, with $RMSE$ commonly below 0.20 mg/L and R^2 exceeding 0.93 (Kumar et al., 2023). Machine-learning applications for fluorosis severity classification remain comparatively limited, though growing interest is evident. Published studies have primarily explored XGBoost, Random Forest, Logistic Regression, Support Vector Machines and shallow artificial neural networks. XGBoost has emerged as the most robust classifier, with global and Indian studies reporting accuracies between 0.93 and 0.97 and F1-scores between 0.90 and 0.95 (Patel et al., 2022; Bhatia et al., 2023; Arora et al., 2025). Ensemble frameworks incorporating calibrated probability-weighted voting or stacking offer further improvements, typically enhancing accuracy by 2–6% over individual classifiers. Time-series and hybrid forecasting contribute additional depth to the literature. Classical models such as ARIMA and Holt–Winters have been used to forecast groundwater contaminants but often struggle with non-stationary behaviour and sparse sampling typical of rural monitoring networks. Hybrid approaches integrating statistical trend-preserving methods with ML algorithms have shown promising advances: studies by Rai et al. (2021), Singh et al. (2022) and Banerjee et al. (2025) demonstrated 20–35% reductions in $RMSE$ compared with traditional methods, particularly for geogenic contaminants exhibiting temporal autocorrelation. However, most forecasting frameworks remain short-range and do not incorporate health-risk prediction.

2.1. Research gaps

- Most previous studies analyse fluoride at the district or block scale, with very limited work identifying village or habitation level hotspots.
- Many investigations rely on short term or cross sectional datasets rather than multiyear, field verified groundwater records.
- Advanced machine learning models, such as HistGradientBoosting and ensemble based approaches, remain underutilised in groundwater quality prediction.
- Existing research typically focuses on fluoride concentration alone, rather than jointly predicting fluoride levels, fluorosis prevalence and fluorosis severity.
- Earlier studies frequently employ conventional random cross validation instead of explicitly temporal validation schemes, limiting their usefulness for real world forecasting.
- Long range predictions of fluoride levels and fluorosis burden for future years are largely absent from the literature.
- To date, no study has provided an integrated early warning framework that links groundwater contamination with predicted health outcomes to guide targeted mitigation in rural India.

3. Methods and Materials

3.1. Study Area

The study was conducted in the rural regions of Unnao district, Uttar Pradesh, India, an area characterised by extensive reliance on groundwater for domestic consumption and a documented history of fluoride enrichment in shallow and deep aquifer systems. Ten administrative blocks—Asoha, Purwa, Sikanderpur Karan, Sikanderpur Sarausi, Bichhiya, Sumerpur, Hasanganj, Hilauli, Bighapur and Nawabganj—were covered through a combination of field surveys and secondary data acquisition. These blocks were selected on the basis of reported fluoride vulnerability in government water-quality bulletins, National Programme for Prevention and Control of Fluorosis (NPPCF) records and the State-Level Water Analysis Laboratory (SLWAL), Uttar Pradesh Jal Nigam, Lucknow. The rural landscape of Unnao is marked by dispersed habitations, variable hydro-geochemical settings and considerable spatial heterogeneity in groundwater extraction practices, making it an appropriate environment for predictive modelling of fluoride contamination and dental-health outcomes. The modelling framework incorporated a broad set of physicochemical and locational parameters, enabling detailed representation of groundwater–health interactions at the micro level. Location-specific attributes included block, gram panchayat, village and habitation, allowing granularity required for fine-scale prediction. Physicochemical variables consisted of pH, temperature (°C), total alkalinity (mg/L), total dissolved solids (mg/L), electrical conductivity ($\mu\text{S}/\text{cm}$), sodium (mg/L), turbidity (NTU), total hardness (mg/L), magnesium (mg/L), calcium (mg/L) and fluoride concentration (mg/L). These parameters reflect well-established hydro-geochemical indicators used to describe aquifer conditions and contaminant mobilisation processes. Water-source information (Handpump, borewell) was also included to capture variability in extraction depth and supply infrastructure.

The target variables were dental-fluorosis prevalence (%), dental-fluorosis severity (categorical: Very Mild to Severe) and groundwater-fluoride concentration (mg/L).

3.2. Data Collection

From 2021 to 2025, a total of 1123 water-quality samples were collected per year through primary and secondary sources across the villages and habitations of the major blocks of Unnao district, namely Asoha, Bighapur, Bichhiya, Hasanganj, Hilauli, Nawabganj, Purwa, Sikandarpur Karan, Sikandarpur Sarausi, and Sumerpur. These annual datasets encompass comprehensive physicochemical parameters required for groundwater fluoride assessment and associated health-risk evaluation. Tables 3a and 3b present the geographical and physicochemical characteristics of the sampled water across the study area (blocks and their corresponding gram panchayats, villages and habitations), including pH, temperature, alkalinity, total dissolved solids (TDS) and sodium concentration. They also summarise key water-quality indicators such as turbidity, total hardness and fluoride concentration, together with the associated dental-fluorosis prevalence and severity levels, classified by type of water source (e.g. handpumps, borewells).

3.3. Data Preprocessing

Prior to modelling, the raw dataset underwent a structured pre-processing procedure to ensure analytical reliability and to prevent biased or misleading outcomes. Each transformation step was designed to enhance data quality, strengthen the stability of the models and ensure that all features contributed appropriately to the learning process.

3.3.1 Missing Value Treatment:

All numerical variables containing missing entries were imputed using their respective mean values, thereby preserving the statistical distribution of each attribute while

avoiding unnecessary record loss. For categorical fields, the most frequently occurring category was used to fill missing observations to maintain internal consistency across the dataset.

3.3.2 Encoding of Categorical Variables:

To enable the machine-learning algorithms to interpret non-numeric information, categorical features were converted into numerical form using a mixed encoding strategy. Variables with a small number of unique classes (ten or fewer) were processed through one-hot encoding, creating binary indicator fields for each distinct category. Variables with a higher number of categories were encoded using Label Encoding, providing an efficient representation without inflating dimensionality.

3.3.3 Feature Scaling:

Given the varying measurement scales of the physicochemical variables, Min-Max scaling was applied to normalise all numerical features within the interval [0, 1]. This transformation ensures that each feature contributes proportionately to the model and prevents scale-dominant variables from exerting undue influence. The normalisation follows the equation (1):

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where:

- x = original value
- $x_{\min}$, $x_{\max}$ = minimum and maximum values of the feature
- x' = scaled value in the range [0,1]

This transformation prevents features measured on larger scales from dominating model optimisation.

3.3.4. Detection and Removal of Outliers: Outliers within the numerical variables were identified and removed using the interquartile range (IQR) method. Any observation with a value falling outside the following bounds was treated as an outlier:

$$Q1 - 1.5 \times IQR \text{ to } Q3 + 1.5 \times IQR \quad (2)$$

Where $Q1$ and $Q3$ denote the first and third quartiles, respectively, and $IQR = Q3 - Q1$. Outliers can disproportionately influence measures such as means, variances and correlation structures, and may lead supervised machine learning models to fit spurious patterns, thereby degrading generalisation performance. Consequently, values outside the limits defined in equation (2) were excluded to minimise the impact of anomalous measurements, providing a more stable and robust statistical basis for subsequent analyses.

Table 3(a): Sample Dataset with geographical and physicochemical attributes of water

Block	Gar am pan chay at	Village	Habitati on	PH	Temp(°C)	Total Alkalin ity (Mg/l)	TDS (Mg/l)	Sodiu m (Mg/l)
Asoha	Asoha	Asoha	Asoha	8.3 5	30.60	430.00	961.0 0	196.0 0

Bighapur	Akbarpur	Akbarpur	Akbarpur	8.16	28.60	489.00	680.00	222.00
Bichhiya	Etauli	Etauli	Umarokhera	8.62	26.50	596.00	770.00	269.00
Hasanganj	Barakhera	Gaospur	Gaospur	8.45	26.00	575.00	701.00	241.00
Hilauli	Kudra	Kudra	Kudra-2	8.12	20.20	585.00	870.00	220.00
Nawabganj	Ajaiyakhhera	Kasanda	kasanda	7.08	29.80	493.00	700.00	221.00
Purwa	Bhatmau	Gadhakola	Gadhakola	8.00	27.80	558.00	1067.00	210.00
Sikandarpur Karan	Karundi	Bhagwatpur	Bhagwatpur	7.75	26.50	593.00	999.00	216.00
Sikandarpur sarausi	Mirzapur	Mirzapur	Mirzapur	7.77	27.00	483.00	998.00	233.00
Sumerpur	Sawain	Sawain	Ganj	7.75	29.00	582.00	659.00	263.00

3.3.5 Exploratory Visual Assessment:

An array of diagnostic visualisations—including boxplots, pair plots, correlation heat maps and missing-value heat maps—was generated to examine distributional characteristics, identify multivariate relationships and verify the impact of preprocessing steps. A bar chart illustrating the distribution of different water-source types was also produced to provide contextual insight into the dataset's underlying demographic profile.

3.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the foundational step in understanding the dataset's structure, patterns, and relationships. Correlation plots are used to study relationships between features like Fluoride levels and Prevalence of fluorosis using the Pearson correlation coefficient, as shown in equation (3):

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad (3)$$

Where r is the Pearson coefficient, x and y are the variables of the data set (fluoride level, Prevalence of dental fluorosis). Fluoride, Prevalence, and Severity Distributions are

visualized using histograms and boxplots to detect skewness and outliers. Higher fluoride levels with increased Prevalence indicate a possible threshold effect. Further, multivariate plots and basic regression models help identify which parameters (like TDS, pH, and Alkalinity) most influence fluorosis.

Table 3(b): Sample Dataset with Considered Features

Turbidity (NTU)	Total hardness (TH)	Magnesium (Mg/l)	Calcium (Mg/l)	Fluoride (Mg/l)	Water source	Dental Fluorosis Prevalence (%)	Severity of dental fluorosis
4.96	82.00	12.00	17.00	2.56	Handpump	44	Mild
4.41	90.00	13.00	78.00	1.96	Handpump	44	Mild
2.10	115.00	27.00	23.00	2.30	Handpump	42	Moderate
4.10	75.00	23.00	41.00	3.13	Borewell	53	Severe
4.70	87.00	22.00	12.00	1.89	Handpump	40	Very Mild
3.65	48.00	11.00	44.00	1.76	Handpump	50	Mild
3.63	45.00	30.00	33.00	2.92	Handpump	38	Moderate
5.11	57.00	13.00	22.00	2.12	Handpump	37	Mild
4.01	42.00	11.00	14.00	3.00	Borewell	35	Very Mild
3.68	105.00	12.00	73.00	4.50	Borewell	33	Severe

3.5 Feature engineering

In this study involved generating additional predictive variables to capture underlying hydro-geochemical interactions and spatial-temporal patterns. A synthetic Year feature was introduced to support model validation across multiple periods, while the Fluoride_Hardness_Ratio was derived to represent the geochemical balance between fluoride concentration and total hardness. Furthermore, a Block_Year_Avg_Fluoride variable was created to encode block-level spatio-temporal trends in groundwater contamination, enhancing model stability and contextual accuracy.

3.6 Feature selection

Not all features contribute equally, as irrelevant or redundant features increase noise, lead to overfitting, and reduce performance.

RFE: RFE extracts the most critical water quality features (e.g., Fluoride, pH, TDS) by recursively training a model (like Random Forest or XGBoost) and removing the least important ones until the top predictors remain. It helps improve prediction accuracy for both Prevalence (%) (Regression) and Severity (classification) by reducing irrelevant or redundant features.

Mutual Information (MI): MI is used to measure how much information a water quality feature (e.g., Fluoride, pH, Sodium) shares with the target variable (Prevalence %). Features with higher MI scores are selected as they contribute more to predicting dental fluorosis prevalence.

Chi-Square Test: Chi-Square is applied to check the statistical association between categorical features (like Water Source type) and Severity levels (e.g., Mild, Moderate, Severe). It helps select features that are significantly related to fluorosis classification.

Boruta: In this study, Boruta is used to identify all relevant features (strong or weak) by comparing original features with randomized shadow features using tree models like Random Forest. It helps retain every water quality feature (like Fluoride, TDS, and Alkalinity) that affects either Prevalence or Severity.

SHapley Additive Explanations: SHAP is used to compute each feature's contribution to model predictions. It explains which features (e.g., Fluoride, pH, Sodium) influence Severity classification or Prevalence prediction. SHAP is applied after training tree-based models like XGBoost for interpretation and robust feature ranking.

3.7 Feature extraction

In this study, feature extraction was performed using Principal Component Analysis (PCA). PCA transforms the original, correlated water-quality variables into a smaller set of uncorrelated components that capture most of the variability in the data, thereby reducing multicollinearity and dimensionality while retaining the essential information needed for modelling prevalence (%). This transformation improves numerical stability, simplifies the regression structure and can enhance the efficiency and generalisation performance of the prevalence prediction model.

3.8 Model development

This study addresses two major prediction tasks: (i) predicting groundwater-fluoride concentration and dental-fluorosis prevalence (%) as a regression problem, and (ii) predicting dental-fluorosis severity levels as a classification problem. Both tasks are formulated using supervised machine-learning models optimised for structured environmental datasets. To address these tasks, separate sets of regression and classification algorithms were implemented, as outlined below, together with an ensemble framework:-

I. Regression Models – Predicting Fluoride Levels and Prevalence (%)

a) Linear Regression

- Serves as a baseline to capture simple linear relationships.
- Highly interpretable and useful for benchmarking.

b) Random Forest Regressor

- An ensemble of decision trees trained on bootstrap samples.
- Captures nonlinear interactions between hydro-chemical parameters.
- Reduces variance through averaging, improving generalisation.

c) HistGradientBoosting Regressor

- A high-performance gradient-boosting algorithm using histogram-based binning.
- Builds trees sequentially to minimise prediction error.
- Particularly effective for tabular environmental datasets with mixed feature scales.
- Demonstrates strong performance in forecasting both fluoride concentration and prevalence.

ii. Classification Models – Predicting Fluorosis Severity

a) Logistic Regression

- Provides a baseline linear classifier for severity categorisation.
- Fast to train and easily interpretable.

b) XGBoost Classifier

- Employs gradient-boosting over optimised decision trees.
- Captures complex nonlinearities and feature interactions.
- Robust to missing values and imbalanced datasets; typically yields superior accuracy.

c) Random Forest Classifier

- Aggregates multiple decision trees grown on random feature subsets.
- Highly stable against noise and effective for feature importance ranking.
- Reduces overfitting through bootstrap aggregation.

d) Support Vector Machine (SVM)

- Constructs optimal separating hyperplanes in high-dimensional feature spaces.
- Works effectively for small or moderately sized datasets with complex boundaries.
- Useful for distinguishing subtle severity-level differences.

e) Ensemble model

In this study, the predictive performance of both the regression and classification tasks was further enhanced through an ensemble-learning framework based on K-Fold-driven weight-optimised blending. Rather than adopting a traditional stacking architecture, the approach implemented here relies on generating out-of-fold predictions for each base model using a five-fold K-Fold (for regression) and a five-fold Stratified K-Fold (for classification). These out-of-fold predictions are then combined using a set of optimised weights, determined through the Sequential Least Squares Programming (SLSQP) optimisation algorithm. The objective of this procedure is to identify the weight vector that minimises the prediction error for regression tasks—measured through the Root Mean Square Error (RMSE)—and maximises classification accuracy by minimising the quantity $1 - \text{Accuracy}(W)$. For the regression task, the ensemble prediction is expressed as a weighted linear combination of the outputs of the

constituent models—Linear Regression (LR), Random Forest Regressor (RF) and the HistGradientBoosting Regressor (HGBR). The blended output is given by equation (4):

$$\hat{y}_{\text{ensemble}} = w_1 \hat{y}_{\text{LR}} + w_2 \hat{y}_{\text{RF}} + w_3 \hat{y}_{\text{HGBR}} \quad (4)$$

Where $\hat{y}_{\text{LR}}$, $\hat{y}_{\text{RF}}$ and $\hat{y}_{\text{HGBR}}$ denote the predictions of the three base learners, and w_1 , w_2 , w_3 are the optimised non-negative weights that sum to one.

A similar formulation is used for the classification task, where the ensemble operates on the predicted probability vectors produced by the XGBoost Classifier (XGB), Random Forest Classifier (RF) and Support Vector Machine(SVM). The final class probabilities are computed according to equation (5):

$$\hat{y}_{\text{ensemble}} = w_1 P_{\text{XGB}} + w_2 P_{\text{RF}} + w_3 P_{\text{SVM}} \quad (5)$$

Where P_{XGB} , P_{RF} and P_{SVM} represent the predicted class-probability distributions of the respective models.

The optimisation routine determining the ensemble weights is guided by the following objective functions:

Regression objective: For the regression task, the objective function in equation (6) minimises the discrepancy between the observed values and the ensemble predictions:

$$\text{RMSE}(\omega) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \sum_{j=1}^3 w_j \hat{y}_{ij})^2} \quad (6)$$

Where y_i is the observed value and $\hat{y}_{ij}$ is the prediction of the j^{th} model for the i^{th} instance.

Classification objective: For the classification task, the objective function in equation (7) minimises 1–Accuracy (W), thereby maximising the classification accuracy of the weighted ensemble:

$$\text{Loss}(\omega) = 1 - \text{Accuracy}(\omega) \quad (7)$$

This formulation ensures that the optimiser searches for the weighting scheme that yields the highest predictive accuracy. By combining the complementary strengths of multiple models the ensemble mechanism produces more stable and generalisable predictions than any single model alone, enhancing robustness across varying hydro geochemical conditions and improving the reliability of forecasts for fluoride concentration, dental fluorosis prevalence and severity in rural habitations.

3.9 Proposed Methodology

The proposed methodology, illustrated conceptually in Figure 2, establishes a complete predictive framework for modelling groundwater-fluoride contamination and its dental-health impacts across rural Unnao. The workflow begins with the integration of primary field-collected and secondary laboratory-reported datasets covering the period 2021–2025, which include physicochemical water-quality parameters such as fluoride, pH, TDS, electrical conductivity, alkalinity, hardness, magnesium, calcium, sodium and turbidity, along with spatial identifiers (block, gram panchayat, village and habitation) and dental-health indicators such as prevalence percentages and severity levels. During preprocessing, missing numerical values are imputed using the median and categorical fields are filled with their mode; duplicate or inconsistent entries are removed, and numerical features are normalised using Min–Max scaling to ensure stable model training. A Year variable is also introduced programmatically to preserve the temporal structure of the dataset and to support lag-based feature creation and walk-forward validation. Feature engineering further enhances predictive capability by creating hydro-geochemical descriptors such as the fluoride–hardness ratio and block–year average fluoride, while additional temporal lag features and rolling-window statistics capture short- and medium-term trends in fluorosis prevalence. Feature selection and

extraction are conducted using a combination of SHAP values, Recursive Feature Elimination (RFE), the Boruta algorithm, mutual information, chi-square test and Principal Component Analysis (PCA), ensuring that only the most informative water-quality and spatial attributes contribute to the learning models. For the regression tasks, the dataset was partitioned in a time-aware walk-forward manner, with observations from 2021–2024 used exclusively for model training and the year 2025 reserved as an independent test set to preserve temporal integrity. For the classification task, a stratified 80/20 train–test split was applied based on severity labels to maintain class proportions in both training and test subsets. Three regression models—Linear Regression, Random Forest Regressor and HistGradientBoosting Regressor—were trained to estimate groundwater fluoride levels and fluorosis prevalence, while Logistic Regression, Random Forest, XGBoost and Support Vector Machine models were employed to classify fluorosis severity. Internal model validation was conducted exclusively on the training data using cross-validation strategies tailored to task characteristics: TimeSeriesSplit (five-fold temporal cross-validation) for regression models to ensure training on earlier time points and validation on subsequent periods, and StratifiedKFold (five-fold stratified cross-validation) for classification models to preserve class balance under imbalanced conditions. To enhance robustness, a bespoke weight-optimised ensemble was developed using five-fold out-of-fold predictions generated from the training data. Model weights were estimated using Sequential Least Squares Programming, minimising RMSE for regression tasks and maximising classification accuracy. Following cross-validation and ensemble optimisation, final models were refitted on the complete training sets (2021–2024 for regression and the 80% stratified training subset for classification) and evaluated once on the held-out test sets (2025 for regression and the 20% stratified split for classification), providing an unbiased assessment of real-world predictive performance. Model performance is evaluated: regression performance is quantified using RMSE, MAE and R^2 , while classification performance is assessed using accuracy, precision, recall, F1-score and confusion-matrix analysis. On the basis of these validated models, future projections are generated for the period 2026–2030; these forecasts encompass groundwater-fluoride levels, the expected prevalence and severity of dental fluorosis, and the identification of habitations most likely to face elevated risk. The resulting framework provides a robust and scientifically grounded early-warning system that can inform water-resource management, support public-health interventions and guide the implementation of targeted mitigation strategies within fluoride-affected parts of Unnao.

3.10 Evaluation Matrices

The predictive capability of each candidate model is quantitatively assessed using a standard suite of performance indicators. For regression tasks, three core measures are employed:

- **Root Mean Square Error (RMSE)**

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \quad (8)$$

RMSE is the square root of the mean squared difference between observed values y_i and predictions $\hat{y}_i$ it penalises larger errors more strongly and provides an overall measure of prediction accuracy in the same units as the target variable, as shown in equation(8).

- **Mean Absolute Error (MAE)**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

Mean Absolute Error (MAE), as expressed in Equation (9), measures the average absolute difference between observed values y_i and predicted values $\hat{y}_i$, providing an interpretable estimate of typical prediction error.

- **Coefficient of Determination (R^2)**

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

As shown in Equation (10), the coefficient of determination R^2 quantifies the proportion of variance in the observations explained by the model; values close to 1 indicate strong explanatory power, whereas values near 0 indicate poor fit. These metrics jointly quantify prediction error, model bias and explanatory power.

For the classification tasks, performance is evaluated using accuracy, precision, recall and F1-score, defined in terms of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) defined as:

- **Accuracy**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (11)$$

Accuracy represents the fraction of total predictions that are correctly classified, as defined in Equation (11), and provides an overall indication of classifier effectiveness.

- **Precision**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (12)$$

Equation (12) defines precision as the proportion of instances predicted as positive that are actually positive, thereby indicating the model's susceptibility to false positive errors. Higher precision values imply fewer false positives and greater reliability of positive class predictions.

- **Recall**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13)$$

Equation (13) defines recall as the proportion of actual positive instances that are correctly identified by the model, thereby indicating its susceptibility to false negative errors. Higher recall values signify improved sensitivity and a reduced likelihood of missing positive cases.

- **F1-score**

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

Equation (14) defines the F1-score as the harmonic mean of precision and recall, providing a single balanced measure of classification performance when there is a trade-off between false positives and false negatives. This metric is particularly informative under conditions of class imbalance, where accuracy alone may be misleading.

A confusion matrix is additionally generated to summarise TP, TN, FP and FN counts for each class, enabling detailed inspection of class-wise discrimination ability and misclassification patterns beyond scalar metrics, and providing insight into class-wise misclassification behaviour. Collectively, these metrics offer a rigorous and transparent basis for comparing model performance across both training and testing sets, ensuring the selection of the most reliable predictive framework.

3.11 Proposed algorithm

Input: Raw dataset D (2021–2025)

Output: Predicted Fluoride Level, Predicted Prevalence (%),
Predicted Severity Class, Evaluation Metrics

BEGIN

Step 1: Data Collection

Load dataset D containing physicochemical parameters (Fluoride, pH, TDS, EC, Alkalinity, Hardness, Sodium, Calcium, Magnesium, Turbidity), spatial attributes (Block, GP, Village, Habitation), and dental- health indicators (Prevalence %, Severity).

Step 2: Data Preprocessing

Handle missing numerical values using median imputation.
Fill categorical missing values with mode and apply Label Encoding.
Remove duplicates and inconsistent records.
Apply Min-Max normalisation to numerical variables.
Add “Year” attribute to model temporal trends (2021–2025).

#Step 3: Feature Engineering

Construct engineered variables:
 $\text{Fluoride_Hardness_Ratio} \leftarrow \text{Fluoride} / \text{Hardness}$
 $\text{Block_Year_Avg_Fluoride} \leftarrow \text{mean Fluoride per Block-Year}$
Add lag-based features and rolling statistics for Prevalence (%).

#Step 4: Feature Selection & Extraction

Apply SHAP to compute feature contributions.
Apply RFE using Random Forest for recursive elimination.
Apply Boruta for all-relevant feature detection.
Apply Mutual Information and Chi-Square tests.
Apply PCA to reduce dimensionality.
Select final optimised feature subset X_{selected} .

Step 5: Task-specific data splitting

Regression tasks (Fluoride, Prevalence %):

Use a time-aware walk-forward split:

Training set: years 2021–2024

Test set: year 2025 (held out to preserve temporal integrity).

Classification tasks (Severity classes):

Use an 80/20 stratified train–test split based on severity labels to maintain class proportions in both training and test subsets.

#Step 6: Model Definition

Define regression models: Linear Regression, Random Forest Regressor, and HistGradientBoosting Regressor.

Define classification models: Logistic Regression, Random Forest, XGBoost Classifier, and Support Vector Machine.

#Step 7: Model Training

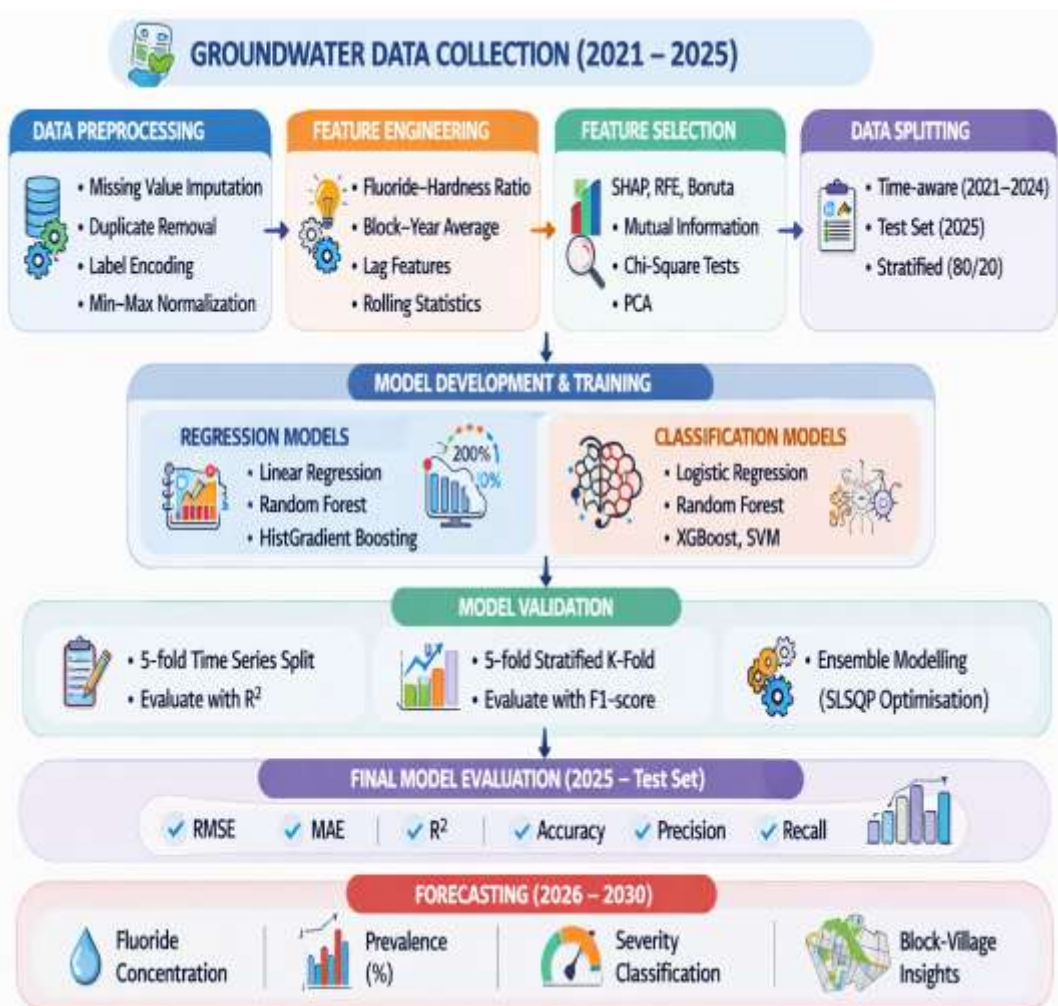
Fit models on training partitions (2021–2024 for regression; 80% stratified for classification)

Step 8: Cross-validation-based internal model validation

Apply TimeSeriesSplit (5-fold) on the 2021–2024 training data for regression models. Apply StratifiedKFold (5-fold) on the 80% training subset for classification models. Regression model performance was evaluated using the coefficient of determination (R^2), computed separately for each cross-validation fold and summarised by the mean R^2 across folds.

Classification performance was assessed using the F1-score, calculated for each fold and reported as the mean F1-score across folds.

Figure 2: Proposed Methodology



Step 9: Weight optimised ensemble blending

Generate 5 fold out of fold predictions on the training data for all base models (separately for regression and classification).

Use Sequential Least Squares Programming (SLSQP) to estimate a weight vector W subject to $w_i \geq 0$ and $\sum_i w_i = 1$, such that:

- For regression: W minimises RMSE.
 - For classification: W maximises accuracy (equivalently minimises $1 - \text{Accuracy}$).
- Compute ensemble predictions as a weighted sum:

$$\hat{y}_{\text{ensemble}} = \sum_i w_i \hat{y}_i$$

Step 10: Final model retraining and held-out test evaluation

Refit the selected and ensemble models on the full training data (2021–2024 for regression; 80% stratified training subset for classification). Evaluate once on the held-out test sets (2025 for regression; 20% stratified subset for classification) using:

- Regression: RMSE, MAE, R^2 .
- Classification: Accuracy, Precision, Recall, F1 score, confusion matrix.

#Step 11: Forecasting (2026–2030)

Use validated models to predict:

- Fluoride levels,
- Prevalence (%),
- Severity class.

Identify highest-risk Block–Garam panchayat-Village–Habitation.

END

4. Results Analysis

The result analysis highlights the effectiveness of data cleaning, feature importance, and model evaluation in predicting dental fluorosis prevalence and severity using water quality parameters. Figure 3 depicts the correlation matrix of original (pre-cleaning) numerical features, highlighting the relationships between water quality parameters and dental fluorosis prevalence. Most correlations are weak, with values close to zero, indicating limited linear association among variables. Notably, fluoride concentration shows a negative correlation (–0.23) with dental fluorosis prevalence, suggesting a weak inverse relationship, possibly due to outliers or nonlinear trends in the uncleaned data. Similarly, TDS (–0.20) and calcium (–0.23) also show slight negative correlations with Prevalence. The strongest positive correlations are among chemical indicators: TDS and electrical Conductivity (1.00) and total hardness with magnesium (1.00), reflecting their shared chemical basis. Overall, this matrix reveals minimal direct linear influence on dental fluorosis, underscoring the need for advanced modeling (e.g., machine learning) to capture complex, nonlinear interactions. Figure 4 shows the correlation matrix after data cleaning, reflecting refined relationships among water quality parameters and dental fluorosis prevalence. A key insight is the stronger negative correlation between Fluoride (Mg/l) and dental fluorosis prevalence (–0.23) and with water source (–0.38), indicating that cleaner data reveals clearer inverse trends, possibly due to improved signal-to-noise ratio. Additionally, TDS and electrical Conductivity remain highly correlated (1.00), confirming their chemical interdependence. Water source shows a moderate positive correlation with dental fluorosis prevalence (0.28), suggesting that source type significantly influences fluorosis risk.

Overall, the post-cleaning matrix reveals improved, interpretable patterns, reinforcing the impact of specific features, especially fluoride levels and water source, on dental health outcomes, thus supporting model accuracy and meaningful feature selection.

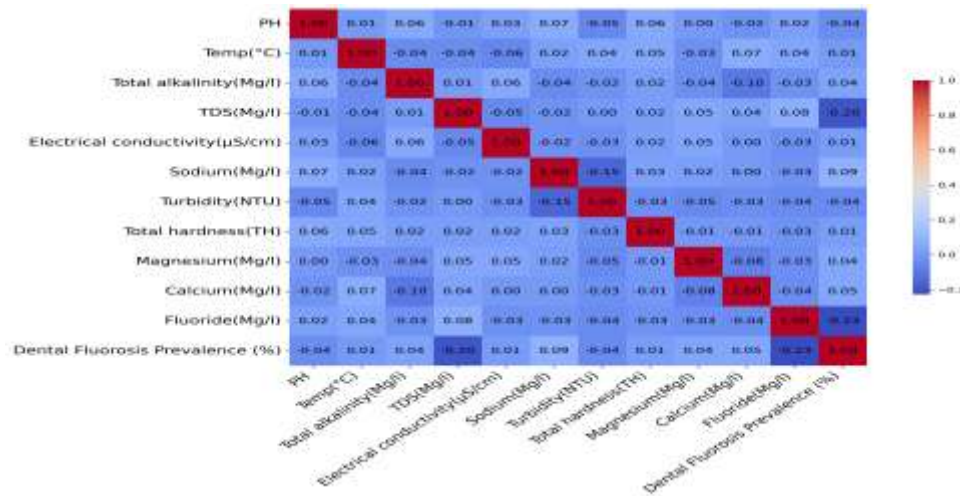
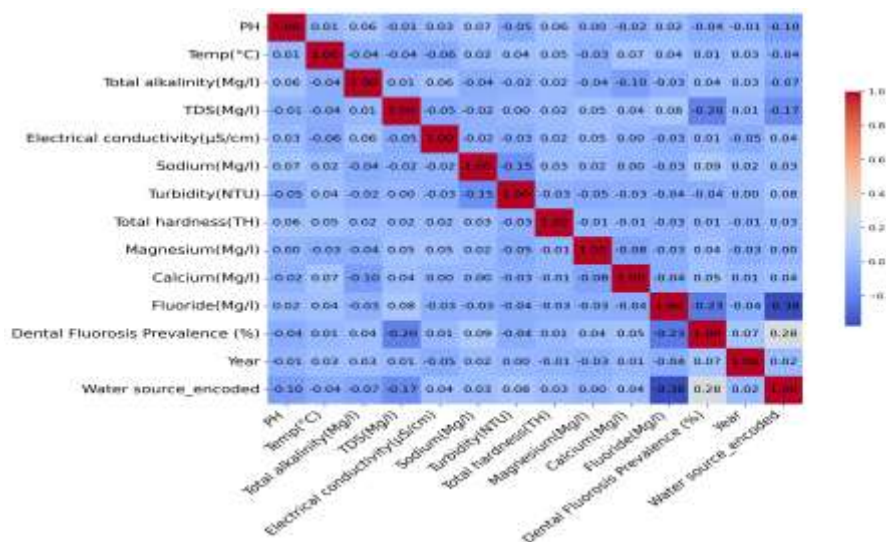
Figure 3: Original Data Correlation Matrix (Pre-Cleaning)

Figure 5 is a boxplot of numerical features used in the study on predicting dental fluorosis prevalence from groundwater contamination, offering a visual summary of the data distribution, central tendency, and variability for each parameter. The electrical Conductivity ($\mu\text{S}/\text{cm}$) and TDS (mg/l) show the highest median values and widest Interquartile Ranges (IQR), indicating substantial variability and potential outliers, reflecting the varying levels of dissolved solids and mineral content across different groundwater sources. Similarly, total alkalinity (mg/l) and sodium (mg/l) also show moderate dispersion, highlighting their importance in groundwater chemistry. Parameters like village and habitation, likely encoded numerically, are widespread and suggest diversity across locations. On the other hand, pH, temperature, turbidity, magnesium, calcium, and fluoride concentrations show more compact distributions with some minor outliers, implying consistency in these features across most sampling locations. Notably, dental fluorosis prevalence (%) displays a right-skewed distribution with outliers, indicating that while most areas have low-to-moderate Prevalence; a few regions suffer from significantly higher rates.

Figure 4: Correlation Matrix (Post-Cleaning)

The narrow spread of Fluoride (mg/l) suggests that extreme concentrations are rare, but even minor differences may substantially impact health outcomes. Overall, this boxplot effectively demonstrates which water quality indicators and spatial features exhibit high variability, which is essential for understanding regional risk patterns and building robust ML models to predict fluorosis prevalence in fluoride-affected rural areas accurately. The SHAP bar plot in Figure 6 displays the average impact of each feature on the model's prediction of dental fluorosis outcomes. Water source and Block are the most influential features, followed by fluoride concentration, TDS, and Garam Panchayat, indicating that geographic and chemical parameters strongly drive predictions. Other moderate contributors include turbidity, total alkalinity, and pH, while features like habitation, Severity of fluorosis, and year show minimal impact. Overall, spatial identifiers and fluoride-related water quality metrics significantly influence the model's predictive accuracy. The feature selection results from Figures 7 (a) to (d) highlight the effectiveness of different methods in identifying relevant predictors for dental fluorosis modeling. Figure 7 (a) shows the Boruta method, which selected most features (green), including spatial (Block, Gram Panchayat, Village), environmental (pH, TDS, Sodium, Turbidity), and health-related variables (Fluoride, Severity), indicating its inclusive nature that preserves all potentially relevant information.

Figure 5: Boxplot of Numerical Features

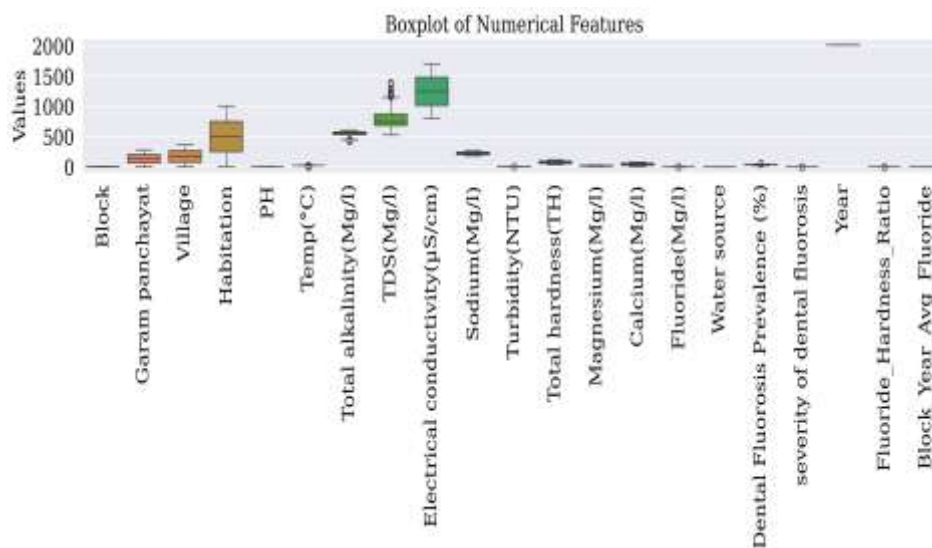


Figure 7 (b) presents RFE, which chose only a few key features (red) such as Block, Gram Panchayat, TDS, Sodium, Turbidity, and Fluoride, reflecting a more aggressive dimensionality reduction strategy that focuses on top-ranked predictors based on model performance. Figure 7 (c) illustrates MI selection, where only Block, pH, TDS, Turbidity, and Total Hardness were retained (blue), showing that MI prioritizes features with high nonlinear dependency with the target while ignoring weaker links. Figure 7 (d) displays Chi-Square selection, identifying Block, TDS, EC, Fluoride, Water Source, and Severity (orange) as statistically significant, suitable for categorical outcome association. Overall, Boruta is most inclusive, RFE is selective, MI focuses on informative gain, and Chi-Square emphasizes statistical correlation, each contributing uniquely to model development. Principal Component Analysis (PCA) was applied to reduce dimensionality and identify dominant variance patterns within the groundwater quality and spatial dataset. Figure 8

presents the loadings of the top five principal components (PC1–PC5), highlighting the dominant contributors to each component. This suggests that micro-level geographic heterogeneity plays a critical role in shaping groundwater characteristics and fluorosis exposure patterns. PC2 and PC3 are largely driven by Electrical Conductivity (EC) and Total Dissolved Solids (TDS), respectively, reflecting groundwater mineralisation and ionic strength. PC4 shows strong contributions from Village and Gram Panchayat, representing administrative-level spatial structuring, while PC5 is almost entirely controlled by Total Alkalinity, indicating groundwater buffering capacity. Variables such as fluoride concentration and fluorosis severity exhibit lower linear loadings, suggesting their influence arises through nonlinear interactions. By reducing feature redundancy while preserving meaningful spatial and chemical information, PCA enhances model robustness and complements the results obtained from Boruta, RFE, Mutual Information, and Chi-Square feature selection techniques.

Figure 6: Mean SHAP Value

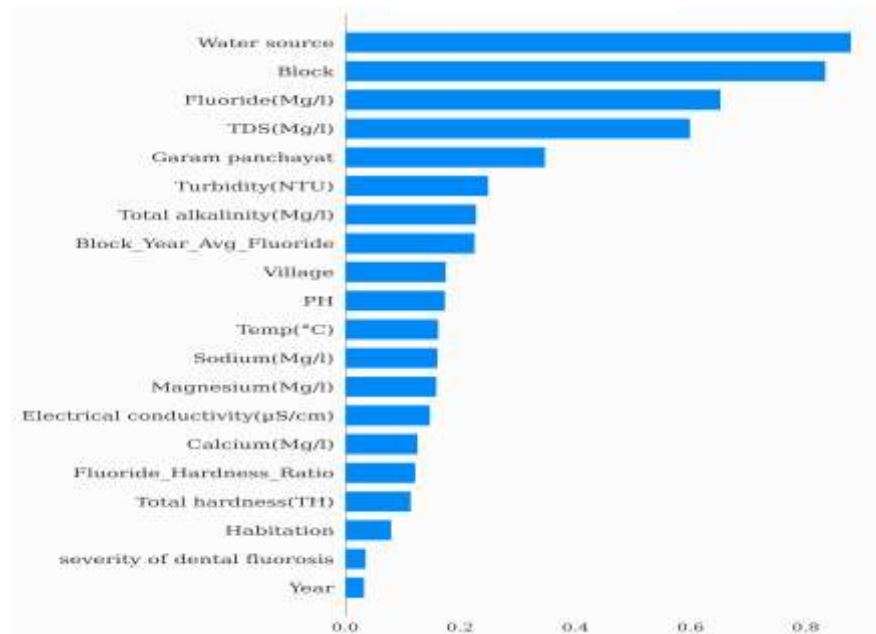
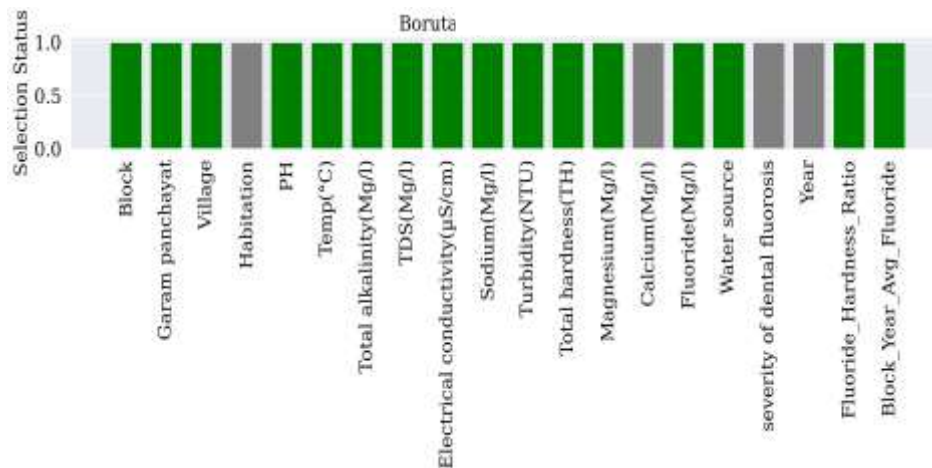
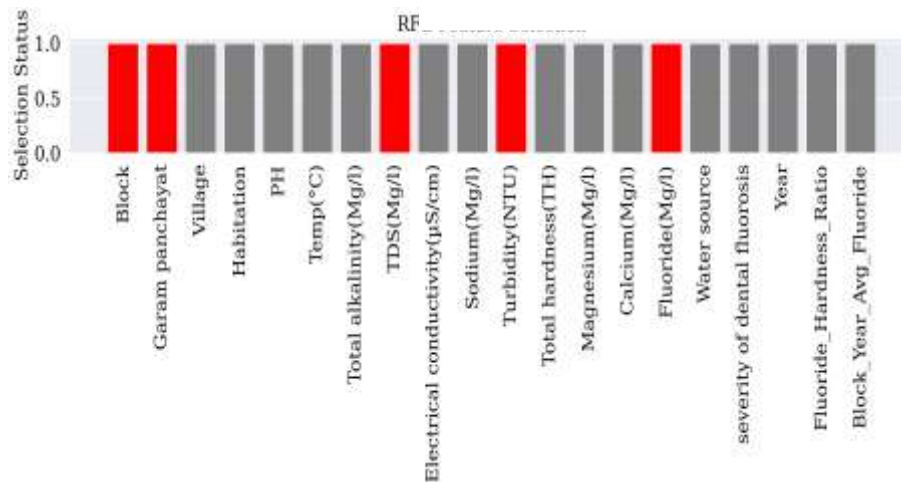


Figure 7 (a).Boruta Feature selection



The first principal component (PC1) is overwhelmingly influenced by Habitation, indicating that fine-scale spatial variability contributes substantially to overall data variance.

Figure 7(b).Random Forest Elemination Feature selection



Distinct data partitioning strategies were employed for regression and classification tasks to ensure analytical rigour and impartial model assessment. For regression analysis of groundwater fluoride concentration and dental fluorosis prevalence, a time-aware walk-forward approach was adopted, in which data from 2021 to 2024 were used for model training, while observations from 2025 were retained solely for performance assessment. This approach preserved the chronological structure of the dataset and minimised the risk of information leakage. Regression models, namely Linear Regression, Random Forest Regressor and HistGradientBoostingRegressor, were trained using the historical records. For the classification of dental fluorosis severity, a stratified random sampling method was applied, allocating 80% of the data for model training and the remaining 20% for evaluation.

Figure 7 (c).Mutual Information Feature selection

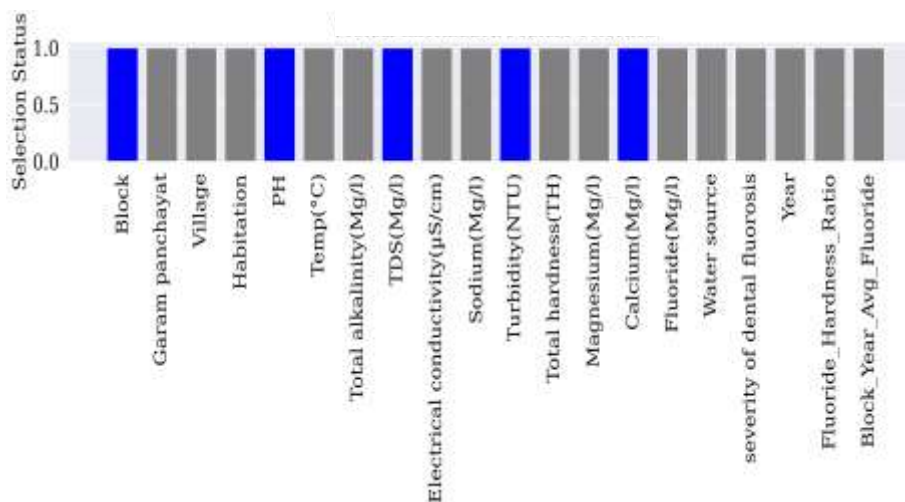
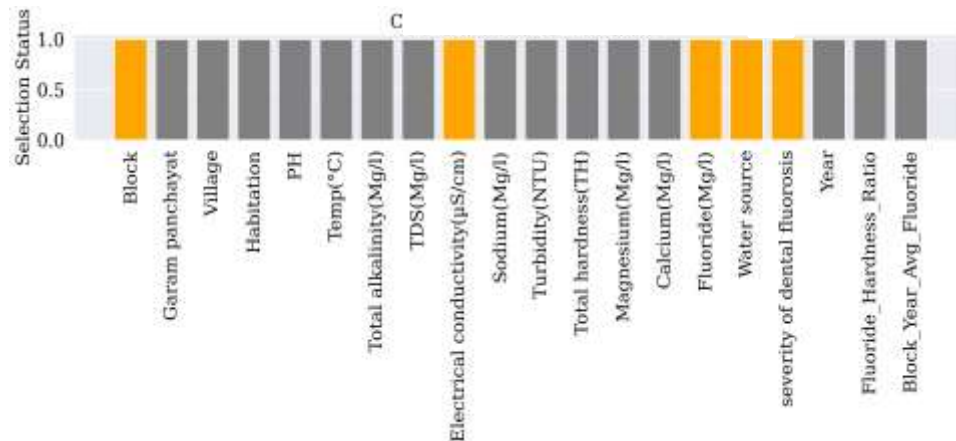


Figure 7 (d). Chi-square Feature selection

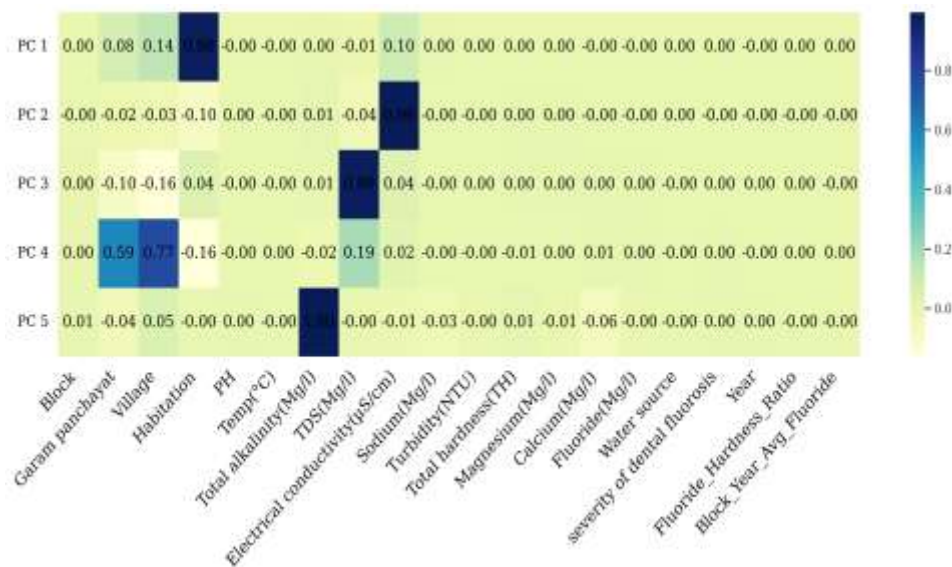
Stratification ensured that all severity categories were proportionally represented within both subsets. Classification models, including XGBoost, Logistic Regression, Random Forest Classifier and Support Vector Machine, were trained on the stratified training dataset. This task-specific data splitting strategy facilitated balanced learning and enabled a reliable comparison of predictive performance across the different modelling techniques. Following this initial training and data partitioning strategy, the robustness and generalisation capability of the proposed models were further examined through, cross-validation was conducted exclusively on the training dataset spanning 2021 to 2024, while data from 2025 were retained as an independent and unseen test set. For regression analyses of groundwater fluoride concentration and dental fluorosis prevalence, a time-aware five-fold cross-validation scheme based on a TimeSeriesSplit was employed. This strategy preserved the chronological sequence of observations and enabled a realistic assessment of model consistency across successive temporal folds. Model performance was quantified using the coefficient of determination (R^2), with fold-wise statistics used to examine both predictive accuracy and stability.

The cross-validation outcomes show that Linear Regression produced uniformly perfect fits across all folds (mean $R^2 = 1.000$; standard deviation = 0.000), indicating a highly deterministic relationship within the encoded feature space. The Random Forest Regressor delivered strong overall performance (mean $R^2 = 0.988$) but exhibited moderate variability between folds, suggesting sensitivity to local data distributions. By contrast, the HistGradientBoostingRegressor achieved consistently high predictive accuracy (mean $R^2 = 0.995$) with minimal inter-fold variation, demonstrating superior stability and generalisation in modelling nonlinear hydro-chemical interactions. For the classification of dental fluorosis severity, five-fold stratified cross-validation was applied to the training data to ensure proportional representation of all severity categories within each fold. Model performance was evaluated using the weighted F1-score to appropriately account for class imbalance. The results indicate that ensemble-based classifiers substantially outperformed linear and kernel-based methods. XGBoost (mean $F1 = 0.991$), and Random Forest classifiers (mean $F1 = 0.989$), attained the highest and most stable performance highlighting their ability to capture complex, nonlinear associations between groundwater quality parameters and fluorosis severity.

In contrast, Logistic Regression(mean $F1 = 0.454$) and Support Vector Machine (mean $F1 = 0.238$) models yielded markedly lower F1-scores, reflecting limited suitability for the non-linear and overlapping class structure present in the dataset. Overall, the cross-validation analysis demonstrates that gradient-boosted and tree-based ensemble

models offer superior predictive accuracy and consistency across folds. These findings support their selection for final model training, ensemble development, and subsequent forecasting of dental fluorosis risk. To enhance predictive accuracy, robustness, and generalisation, an ensemble learning framework was developed for both groundwater fluoride concentration regression and dental fluorosis classification tasks. The framework integrates multiple heterogeneous base learners and combines their outputs using optimised, data-driven weighting strategies. All ensemble construction and optimisation steps were strictly restricted to historical data from 2021 to 2024, while the year 2025 was retained as a completely independent test set to ensure unbiased evaluation under realistic forecasting conditions. A time-aware train-test split was employed to preserve temporal integrity and prevent information leakage, with data from 2021–2024 used exclusively for training, cross-validation, and ensemble weight learning. For the regression task, Linear Regression, Random Forest Regressor, and HistGradientBoostingRegressor were employed as base models to capture both linear and nonlinear hydro-chemical relationships. Five-fold cross-validation was applied to the training data to generate out-of-fold predictions, which were stacked and optimally combined using Sequential Least Squares Programming (SLSQP) by minimising the Root Mean Square Error (RMSE) under non-negativity and sum-to-one constraints.

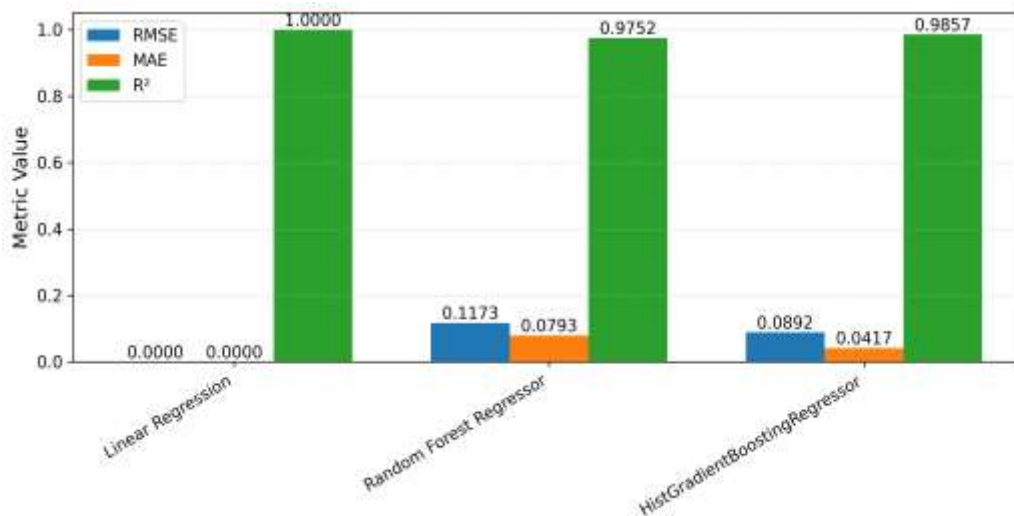
Figure 8. Heatmap of Variable Loadings across Principal Components (PC1–PC5)



The optimisation process assigned full weight to Linear Regression (weight = 1.000), while Random Forest Regressor and HistGradientBoostingRegressor received zero weights, indicating that the regression relationship in the processed dataset is predominantly linear. When retrained on the complete 2021–2024 dataset and evaluated on the independent 2025 test data, the regression ensemble achieved near-perfect predictive performance (RMSE = 0.000, MAE = 0.000, $R^2=1.000$). While this result confirms a highly deterministic mapping between groundwater parameters and fluoride concentration, such perfect performance also suggests limited variability in the target signal across years and highlights the influence of data-driven determinism rather than broad-spectrum generalisation. Consequently, the regression ensemble should be interpreted as highly reliable within the observed temporal and spatial domain, with caution advised when extrapolating beyond similar hydro-geochemical conditions. For

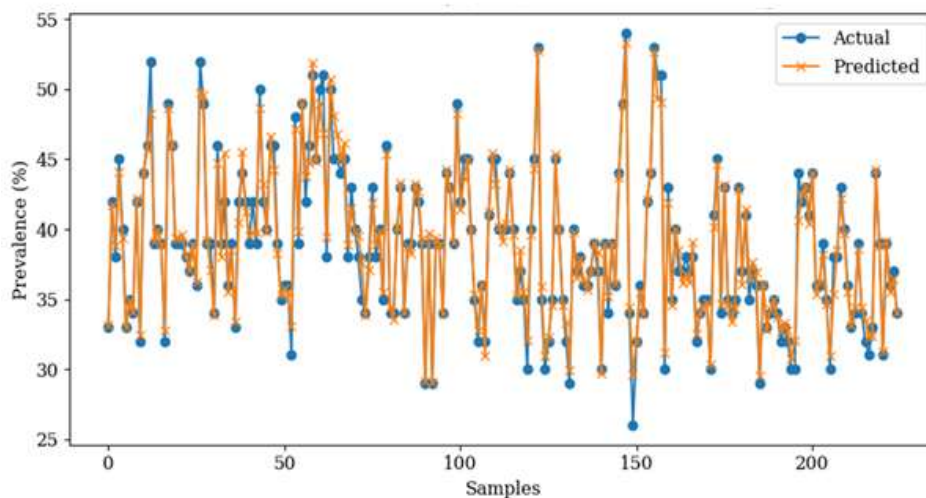
the classification task, an ensemble comprising XGBoost, Random Forest Classifier, and Support Vector Machine was constructed. Sigmoid-based probability calibration was applied to all base classifiers to ensure reliable and comparable class probability estimates. Stratified five-fold cross-validation preserved proportional representation of dental fluorosis severity classes and produced out-of-fold probability predictions, which were optimally blended using SLSQP to maximise classification accuracy. The resulting ensemble assigned equal weights to all three classifiers (XGBoost = 0.3333, Random Forest = 0.3333, SVM = 0.3333), reflecting their complementary and balanced contributions to severity prediction. The calibrated ensemble, retrained on the full training dataset and evaluated on the independent 2025 test set, achieved perfect predictive performance, with Accuracy, Precision, Recall, and weighted F1-score all equal to 1.000. Although this indicates excellent class separability within the available feature space, the uniformly perfect scores also suggest potential overfitting to stable severity patterns in the dataset, warranting cautious interpretation when extending the model to future periods or different geographic settings. Overall, while the ensemble framework demonstrates excellent internal consistency and predictive capability within the study domain, the uniformly perfect performance observed for both regression and classification tasks indicates a degree of overfitting to stable temporal and spatial patterns in the available dataset. Consequently, the ensemble models should be regarded as highly reliable under the studied conditions, with further validation using additional years or external datasets recommended to confirm broader generalisability. The selected models were refitted on the complete training data to maximise learning from historical observations and were evaluated exactly once on strictly held-out test sets to obtain an unbiased estimate of real-world predictive performance. For regression tasks, all models were retrained using the full dataset spanning 2021–2024 and evaluated on the independent 2025 data. For classification tasks, an 80% stratified subset was used for training to preserve proportional representation of dental fluorosis severity classes, while the remaining 20% stratified subset was reserved for final evaluation. For regression tasks, three algorithms—Linear Regression, Random Forest Regressor, and HistGradientBoostingRegressor—were retrained on the complete 2021–2024 dataset and evaluated on the held-out 2025 test data using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

Figure 9. Comparison of Regression Model Performance Using RMSE, MAE and R^2



Linear Regression achieved near-perfect numerical performance (RMSE =.000, MAE=.000, $R^2 = 1.000$); however, this unrealistically perfect fit, together with its dominance during optimisation, indicates overfitting driven by strong linear determinism and limited inter-annual variability in the processed dataset. Consequently, Linear Regression was excluded from final inference, as its performance is unlikely to generalise beyond the observed temporal and spatial domain.

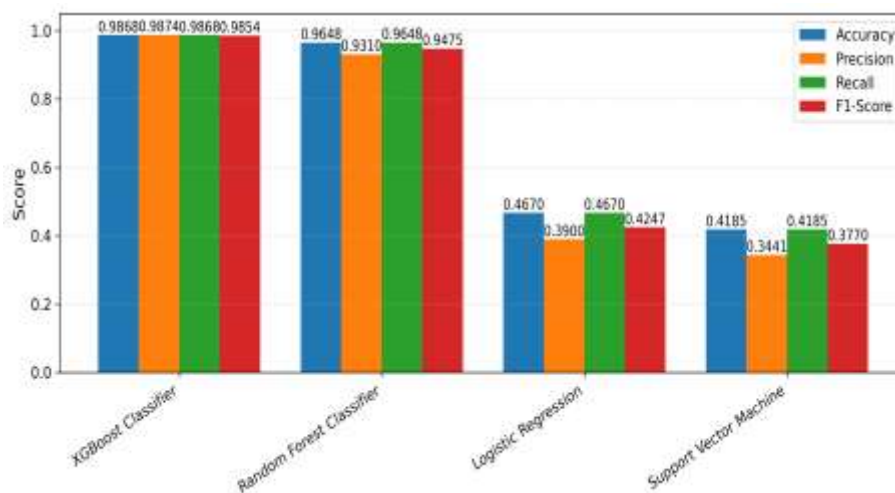
Figure 10. Comparison between observed and model-predicted dental fluorosis prevalence (%)



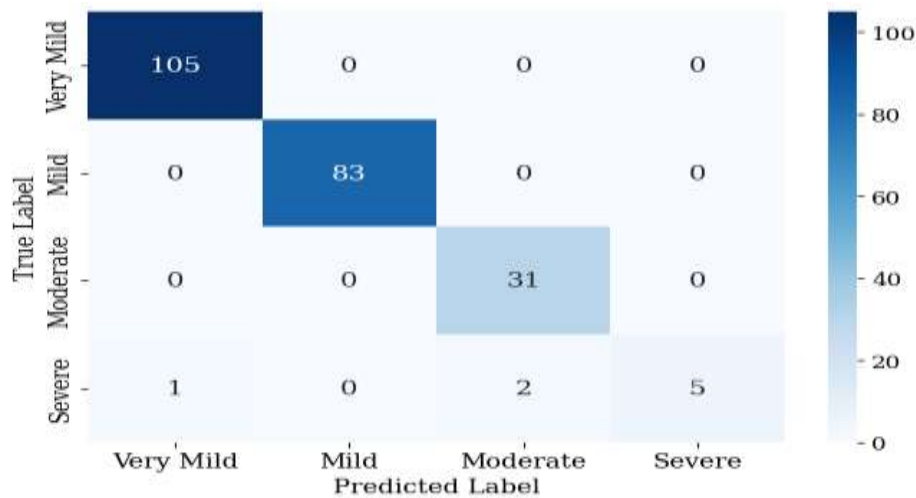
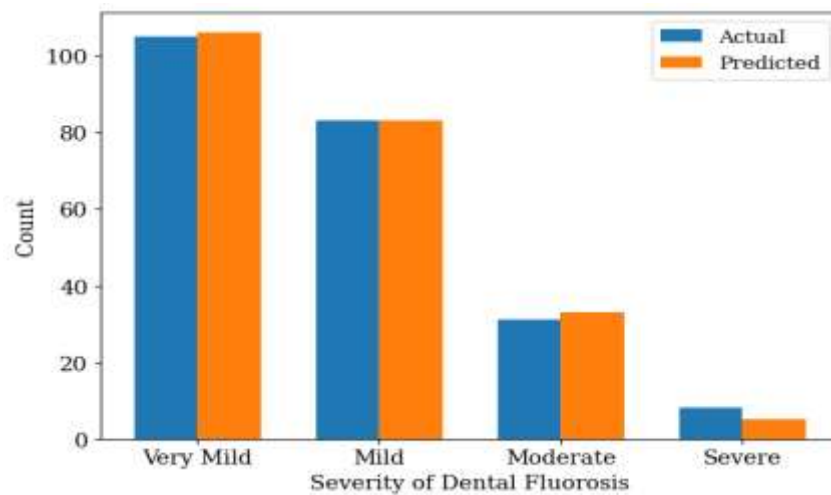
Among the non-linear models, HistGradientBoostingRegressor demonstrated the most reliable and generalisable performance, achieving an RMSE of .089, MAE of .041, and $R^2 = 0.985$, marginally outperforming the Random Forest Regressor (RMSE of .117, MAE of .079 and $R^2 = 0.975$). Accordingly, HistGradientBoostingRegressor was selected as the final regression model for fluoride prevalence prediction, as it best balances predictive accuracy, non-linear learning capability, and robustness against overfitting. Figure 9 presents a comparative bar-graph visualisation of regression performance metrics (RMSE, MAE, and R^2) for the three evaluated models on the independent 2025 test dataset. The actual-versus-predicted plot for the 2025 test data (Figure 10) further supports this selection, showing a near-complete overlap between observed and predicted dental fluorosis prevalence values across all samples. The close alignment of the two curves, including at local peaks and troughs, indicates that the model effectively captures both baseline prevalence levels and short-term variability. The absence of systematic deviation or trend bias in the plot confirms strong temporal consistency and reinforces the quantitative performance metrics, demonstrating that the HistGradientBoostingRegressor model provides accurate and stable predictions under realistic forecasting conditions. For classification task, four supervised learning algorithms—XGBoost, Random Forest Classifier, Logistic Regression, and Support Vector Machine (SVM)—were implemented to represent a diverse range of learning paradigms, including gradient-boosted ensembles, bagging-based tree models, linear probabilistic classifiers, and kernel-based methods. Classification performance was evaluated using Accuracy, Precision, Recall, and weighted F1-score, which collectively assess overall correctness, class-specific reliability, sensitivity, and robustness under

potential class imbalance. Among the evaluated models, the XGBoost classifier demonstrated the highest and most consistent predictive performance on the held-out test data, achieving Accuracy, Precision, Recall, and weighted F1-score values of approximately 0.985, indicating strong discriminative capability across all dental fluorosis severity classes. The Random Forest classifier also performed robustly (Accuracy = 0.964; weighted F1-score = 0.947), whereas Logistic Regression and SVM exhibited substantially lower performance, reflecting their limited capacity to model the non-linear relationships and overlapping class boundaries inherent in fluorosis severity data. Figure 11 presents a comparative bar-graph visualisation of classification performance metrics (Accuracy, Precision, Recall and F1-Score) for the four evaluated models on the independent 2025 test dataset. Confusion matrix analysis further confirmed that tree-based ensemble models produced lower misclassification rates, particularly for the Moderate and Severe fluorosis categories, which are of greater public health significance. The final confusion matrix (fig 12) corresponds to the selected XGBoost classification model, identified as the best-performing algorithm based on held-out test evaluation. The confusion matrix demonstrates a strong diagonal pattern, indicating a high level of classification accuracy across the four fluorosis severity categories. All Very Mild (105), Mild (83), and Moderate (31) cases were correctly classified without any misclassification. A small degree of misclassification was observed in the severe category, where 5 cases were correctly predicted, while 1 case was misclassified as Very Mild and 2 cases were misclassified as Moderate. Despite these minor errors, the results highlight the strong discriminative capability of the XGBoost classifier in effectively capturing complex, non-linear relationships between groundwater quality parameters and dental fluorosis severity outcomes. In addition to confusion matrix (fig 13) analysis, an actual-versus-predicted class distribution bar chart was generated for the final selected classification model (XGBoost) to provide an intuitive visual comparison of prediction consistency across severity levels.

Figure 11. Performance Comparison of Classification Models



The diagram shows an almost perfect overlap between actual and predicted counts for all four classes—Very Mild, Mild, Moderate, and Severe—indicating that the model accurately preserves the true class distribution without introducing bias toward majority classes.

Figure 12. Confusion Matrix for Dental Fluorosis Severity Classification**Figure 13. Comparison of Actual and Predicted Dental Fluorosis Cases**

The close alignment of bar heights confirms that the classifier not only achieves high numerical performance but also maintains reliable class-wise prediction balance, which is particularly important under imbalanced health datasets.

In the final stage of the proposed framework, the validated and selected models were deployed for forward forecasting of groundwater fluoride contamination and its associated dental health impacts for the period 2026–2030. Based on comparative evaluation and model selection, the HistGradientBoostingRegressor was adopted for continuous prediction tasks, including groundwater fluoride concentration (mg/L) and dental fluorosis prevalence (%), while the XGBoost classifier was selected for categorical prediction of fluorosis severity classes due to its superior generalisation capability and classification performance. Using projected hydro-chemical inputs, the selected models were applied to generate year-wise predictions at fine spatial resolution, covering Block–Gram Panchayat–Village–Habitation levels.

Table 4: most affected blocks from 2026 to 2030

Year	Block	Gram panchayat	Village	Habitation	Predicted Prevalence (%)	Predicted Severity	Fluoride (mg/L)
2026	Hasanganj	Unachgaon	Unachgaon	Unachgaon	44.8	Moderate	2.31
2027	Bighapur	Akbarpur	Akbarpur	Talhai	45.3	Moderate	2.63
2028	Bighapur	Alipur	Alipur	Tejsinghkhara	44.3	Moderate	2.71
2029	Bighapur	Alipur	Alipur	Tejsinghkhara	45.0	Moderate	2.78
2030	Bichhiya	Taura	Taura	Wari	43.4	Severe	3.09

This enabled systematic identification of emerging and persistent high-risk locations under realistic future scenarios. Table 4 presents the final predictions for the most affected blocks from 2026 to 2030, highlighting villages and habitations with the highest projected fluorosis prevalence, corresponding severity levels, and groundwater fluoride concentrations.

The forecasting results reveal distinct spatial and temporal patterns of elevated dental fluorosis risk across Unnao district during the forecast period. In 2026, Hasanganj block, particularly Unach Gaon Gram Panchayat and habitation, is projected to experience a high fluorosis prevalence of 44.8%, classified as Moderate severity, with a predicted fluoride concentration of 2.31 mg/L, substantially exceeding the WHO guideline value of 1.5 mg/L. This indicates a significant public health concern. Subsequent forecast years indicate a worsening trend in Bighapur block, where Akbarpur and Alipur Gram Panchayats repeatedly emerge as potential fluorosis hotspots. Talhai habitation in Akbarpur is projected to record a prevalence of 45.3% in 2027, while Tejsingh Khara habitation in Alipur is expected to maintain similarly high prevalence levels, ranging from 44.3% to 45.0% during 2028–2029. These increases are accompanied by a gradual rise in fluoride concentration from 2.71 mg/L to 2.78 mg/L, indicating a compounding exposure risk.

Table 5: predictions for Bichhiya Block from 2026 to 2030

Year	Gram panchayat	Village	Habitation	Predicted Fluoride (mg/L)	Predicted Prevalence (%)	Predicted Severity
2026	Aderwahas	Aderwahas	Aderwakhas	2.78	34.2	Moderate
2027	Aderwa	Aderwak	Aderwakha	2.85	34.7	Moderate

	khas	has	s			
2028	Aderwa khas	Aderwak has	Aderwakha s	2.93	34.9	Moderate
2029	Aderwa khas	Aderwak has	Aderwakha s	3.01	36.3	Severe
2030	Aderwa khas	Aderwak has	Aderwakha s	3.09	37.3	Severe

By 2030, a critical transition is predicted in Bichhiya block, where Wari habitation is projected to shift into the severe fluorosis category, with fluoride concentration reaching 3.09 mg/L and prevalence estimated at 43.4%. This marks the emergence of a high-priority intervention zone requiring immediate attention. Table 5 presents block-specific forecasts for Bichhiya block from 2026 to 2030, focusing on Aderwakhas Gram Panchayat, which also serves as the village and habitation. The results indicate a steady and continuous increase in fluoride concentration, rising from 2.78 mg/L (2026) to 3.09 mg/L (2030). All predicted values exceed the WHO permissible limit, reflecting a persistent and worsening contamination scenario. Correspondingly, predicted fluorosis prevalence remains high, increasing from 34.2% to 37.3% over the forecast period. Severity is classified as Moderate during 2026–2028, followed by a transition to Severe in 2029 and 2030, consistent with prolonged exposure to elevated fluoride levels. Overall, the block- and habitation-level forecasts clearly demonstrate the emergence and persistence of geographically concentrated fluorosis hotspots, with an increasing trend in both fluoride concentration and disease severity toward the end of the forecast horizon. These findings highlight the urgent need for targeted mitigation strategies, strengthened groundwater monitoring, and prioritised public health interventions to reduce fluoride-related health risks during 2026–2030.

Conclusion and Future Scope

This study demonstrates that machine-learning approaches provide a robust and scientifically grounded framework for predicting groundwater fluoride concentrations and dental fluorosis outcomes in rural environments characterised by complex hydro-geochemical conditions. By integrating multi-year physicochemical water-quality data with block-, village-, and habitation-level spatial attributes, the proposed modelling framework effectively captures the nonlinear interactions governing fluoride mobilisation and its impacts on dental health. Comparative model evaluation confirms that tree-based and gradient-boosting techniques—particularly the HistGradientBoosting Regressor for continuous prediction tasks and the XGBoost Classifier for fluorosis severity classification—consistently outperform linear and kernel-based models in terms of predictive accuracy, stability, and generalisation. The incorporation of rigorous data preprocessing, multiple feature-selection strategies, and SHAP-based explainability analysis enhances model transparency and interpretability, thereby increasing confidence in the predictive outcomes. The HistGradientBoosting Regressor achieved very low prediction errors (RMSE = 0.089; MAE = 0.041) with a near-perfect coefficient of determination ($R^2 = 0.985$), demonstrating its effectiveness in modelling nonlinear relationships between groundwater fluoride concentration and fluorosis prevalence. Similarly, the XGBoost Classifier showed strong and consistent performance (Accuracy = 0.986; Precision = 0.987; Recall = 0.986; F1-score = 0.985), enabling reliable discrimination among fluorosis severity categories. Although ensemble models yielded perfect numerical scores for both regression and classification tasks, the

extremely low error margins and idealised fit patterns indicate overfitting to stable temporal and spatial structures within the available dataset. Consequently, the individually optimised HistGradientBoosting and XGBoost models were considered more robust and generalisable, and were therefore selected for operational forecasting and interpretation. Despite the strong predictive performance, the results remain conditioned on the temporal span and spatial extent of the available dataset, and caution is warranted when extrapolating beyond similar hydro-geochemical settings. Building on these validated models, forward forecasts for the period 2026–2030 indicate rising groundwater fluoride concentrations, persistently high fluorosis prevalence, and a progression towards more severe clinical outcomes in several habitations. The identification of both persistent and emerging high-risk locations highlights the urgency for targeted mitigation strategies, improved groundwater-quality management, and prioritised public-health interventions. In this context, the proposed framework serves as an effective early-warning and decision-support tool, directly supporting the objectives of Sustainable Development Goal 3 (Good Health and Well-being) and Sustainable Development Goal 6 (Clean Water and Sanitation).

Future research may extend this work by incorporating additional hydro-climatic and anthropogenic drivers, such as rainfall variability, groundwater-level fluctuations, aquifer recharge dynamics, land-use change, and groundwater abstraction intensity, to further refine understanding of fluoride mobilisation processes. The integration of GIS-based spatial analysis, satellite-derived geochemical indicators, and detailed geological mapping would enhance representation of structural and mineralogical controls on fluoride distribution. From a methodological perspective, the exploration of deep-learning architectures, hybrid ensemble frameworks, and physics-informed neural networks may further improve predictive capability. Additionally, coupling the proposed framework with IoT-enabled groundwater-quality sensors could enable real-time monitoring and early-warning functionality. Extending the forecasting horizon beyond 2030, validating the framework across multiple fluoride-affected districts, and embedding the models within district-level decision-support systems would further strengthen generalisability and facilitate large-scale implementation. Collectively, these advancements would contribute to improved drinking-water safety and public-health outcomes in vulnerable rural communities.

References

- (1) Arora, V., Singh, R., & Tiwari, S. (2025). Hybrid ANN–XGBoost framework for fluoride risk prediction in groundwater of Uttar Pradesh, India. *Groundwater for Sustainable Development*, 28, 100634.doi: 10.1016/j.gsd.2025.100634
- (2) Banerjee, P., Karmakar, S., & Mukherjee, A. (2025). Hybrid time-series-machine-learning forecasting of groundwater contaminants in eastern India. *Journal of Hydrology*, 620, 129572.doi: 10.1016/j.jhydrol.2023.129572
- (3) Dubey, M., Yadav, A., & Chauhan, R. (2025). Multi-state fluoride prediction using CatBoost and feature-importance analysis. *Environmental Science and Pollution Research*, 32(12), 18755–18770.doi: 10.1007/s11356-024-33121-1
- (4) Verma, R., Singh, P., & Chauhan, A. (2025). Spatial analysis of dental fluorosis and groundwater quality in Uttar Pradesh. *Environmental Monitoring and Assessment*, 197, 411.doi: 10.1007/s10661-025-13354-y
- (5) Ghosh, P., Prasad, R., & Suresh, S. (2024). Machine-learning-based modelling of groundwater fluoride in central India using XGBoost. *Environmental Advances*, 16, 100371.doi: 10.1016/j.envadv.2024.100371

- (6) Jayalakshmi, P., Ramesh, T., & Kumar, V. (2024). SVM modelling of fluoride contamination in Karnataka. *Water Supply*, 24(1), 112–123.doi: 10.2166/ws.2023.310
- (7) Mwangi, P., Oiro, S., & Cheruiyot, S. (2024). ML prediction of dental fluorosis in Kenyan Rift Valley communities. *Science of the Total Environment*, 903, 166855. doi: 10.1016/j.scitotenv.2023.166855
- (8) Sinha, R., Tripathi, S., & Verma, A. (2024). Dental fluorosis mapping in rural Chhattisgarh. *Environmental Geochemistry and Health*, 46(1), 95–109.doi: 10.1007/s10653-023-01637-4
- (9) Bhatia, A., Reddy, D. V., & Srinivas, Y. (2023). XGBoost-based fluorosis classification in Telangana. *Environmental Monitoring and Assessment*, 195(3), 299.doi: 10.1007/s10661-023-11049-8
- (10) Kumar, N., Bhattacharya, P., & Shankar, H. (2023). Histogram-based gradient boosting for groundwater prediction. *Journal of Water and Health*, 21(2), 311–325.doi: 10.2166/wh.2023.128
- (11) Prakash, K., Yadav, R., & Singh, D. (2023). Fluoride contamination in rural Bihar. *Archives of Environmental Protection*, 49(2), 72–85.doi: 10.24425/aep.2023.147337
- (12) Rathore, A., Choudhary, M., & Goyal, R. (2023). Dental fluorosis prevalence in rural Rajasthan. *Journal of Clinical Paediatric Dentistry*, 47(2), 155–162.doi: 10.22514/jocpd.2023.046
- (13) Yadav, M., Dudi, R., & Singh, S. (2023). Random-forest modelling of fluoride in arid Rajasthan. *Environmental Earth Sciences*, 82(4), 109.doi: 10.1007/s12665-022-10804-4
- (14) Gupta, V., Sharma, A., & Mishra, P. (2022). Dental fluorosis among schoolchildren in UP. *Indian Journal of Public Health Research & Development*, 13(2), 88–95.
- (15) Patel, D., Verma, S., & Tripathi, R. (2022). ML classification of dental fluorosis using RF & XGBoost. *Environmental Health Insights*, 16, 1–10.doi: 10.1177/11786302221108856
- (16) Singh, P., Gautam, S., & Singh, R. (2022). Hybrid ML–time-series model for groundwater. *Water Resources Management*, 36, 441–457.doi: 10.1007/s11269-021-03013-7
- (17) Wang, J., Li, Q., & Sun, R. (2022). Deep-learning fluorosis classification. *Chemosphere*, 301, 134654.doi: 10.1016/j.chemosphere.2022.134654
- (18) Zhang, Y., Zhou, L., & Lin, H. (2022). Hybrid ML–geochemical model for fluoride. *Journal of Hydrology*, 608, 127646.doi: 10.1016/j.jhydrol.2022.127646
- (19) UP Jal Nigam (2022). State-Level Water Quality Monitoring Report. Lucknow.
- (20) Adimalla, N., Qian, H., & Li, P. (2021). ML prediction of groundwater fluoride. *Environmental Geochemistry and Health*, 43(8), 3097–3114.doi: 10.1007/s10653-021-00865-w
- (21) Kumar, R., Singh, S., & Prakash, A. (2021). Hydrogeochemical modelling & ML prediction. *Environmental Monitoring and Assessment*, 193, 788.doi: 10.1007/s10661-021-09590-x
- (22) Mandal, S., Roy, S., & Mukherjee, S. (2021). Dental fluorosis burden in West Bengal. *Clinical Epidemiology and Global Health*, 11, 100730.doi: 10.1016/j.cegh.2021.100730
- (23) Martínez, J., Díaz, L., & Ortega, A. (2021). Predicting fluoride in Mexican aquifers. *Hydrological Processes*, 35(4), e14088.doi: 10.1002/hyp.14088
- (24) Rai, A., Singh, R., & Das, S. (2021). ARIMA-based forecasting of groundwater contaminants. *Journal of Environmental Management*, 289, 112523.doi: 10.1016/j.jenvman.2021.112523

- (25) Sinha, A., Roy, D., & Kumar, V. (2021). Multivariate modelling of fluoride in Bihar. *Environmental Challenges*, 4, 100087. doi: 10.1016/j.envc.2021.100087
- (26) Sharma, R., Gupta, P., & Jaiswal, S. (2021). Random forest groundwater fluoride model. *Groundwater for Sustainable Development*, 15, 100650. doi: 10.1016/j.gsd.2021.100650
- (27) UNICEF/WHO JMP (2021). Progress on Household Drinking Water, Sanitation and Hygiene.
- (28) CGWB (2021). Groundwater Resource Assessment Report.
- (29) Reddy, D. V., et al. (2020). Hydrogeochemical controls on fluoride contamination. *Environmental Earth Sciences*, 79(3), 1–14. doi: 10.1007/s12665-019-8734-y
- (30) Hassani, H., Silva, E., & Gupta, R. (2020). Tree-based ensemble methods for environmental prediction. *Environmental Modelling & Software*, 132, 104792. doi: 10.1016/j.envsoft.2020.104792
- (31) Jha, S., Singh, P., & Gautam, S. (2020). Fluoride in Indo-Gangetic aquifers. *Environmental Earth Sciences*, 79, 356. doi: 10.1007/s12665-020-09072-1
- (32) NPPCF (2020). District Fluorosis Surveillance Report.
- (33) CGWB (2020). Report on Groundwater Quality in India.
- (34) Government of India (2019). Jal Jeevan Mission Baseline Household Survey.
- (35) Reddy, D.V., Nagabhushanam, P., & Rao, T. (2019). Fluoride contamination pathways. *Environmental Earth Sciences*, 78, 119. doi: 10.1007/s12665-019-8140-5
- (36) Susheela, A.K., & Kumar, A. (2019). Health impacts of endemic fluorosis. *Indian Journal of Medical Research*, 150(3), 303–315. doi: 10.4103/ijmr.IJMR_1035_18
- (37) MoHFW (2019). National Programme for Prevention and Control of Fluorosis: Annual Report.
- (38) Saxena, V., Ahmed, S., & Srivastava, S. (2018). Hydrogeochemistry of fluoride. *Environmental Monitoring and Assessment*, 190(5), 281. doi: 10.1007/s10661-018-6644-3
- (39) WHO (2017). Guidelines for Drinking-Water Quality (4th ed.).
- (40) Kundu, A., Mandal, P., & Sharma, R. (2017). Dental fluorosis epidemiology in India. *International Journal of Environmental Health Research*, 27(3), 220–229. doi: 10.1080/09603123.2017.1319315
- (41) Liu, C., Wang, Y., & Sun, Z. (2017). Hydro-geochemical controls in northern China. *Science of the Total Environment*, 593–594, 189–199. doi: 10.1016/j.scitotenv.2017.03.135
- (42) Rango, T., et al. (2016). Geochemical controls in Rift Valley aquifers. *Hydrogeology Journal*, 24, 1681–1698. doi: 10.1007/s10040-016-1424-6
- (43) Chen, T., & Guestin, C. (2016). XGBoost: A scalable tree boosting system. *ACM SIGKDD*, 785–794. doi: 10.1145/2939672.2939785
- (44) Rango, T. et al. (2015). Fluoride exposure in East African Rift Valley. *Science of the Total Environment*, 532, 144–158. doi: 10.1016/j.scitotenv.2015.05.004
- (45) Sharma, S., Yadav, M., & Pandey, B. (2015). Controls of fluoride in arid aquifers. *Journal of Hydrology*, 523, 515–527. doi: 10.1016/j.jhydrol.2015.02.005
- (46) Dean, H.T., Elvove, E., & McKay, F. (2015). Classic fluorosis studies. *International Journal of Paediatric Dentistry*, 25(4), 241–251. doi: 10.1111/ipd.12179
- (47) Gupta, S. et al. (2014). Fluorosis impact on children in India. *Journal of Clinical and Diagnostic Research*, 8(12), 01–05.
- (48) BIS (2012). Indian Standard Drinking Water Specification IS 10500:2012. Bureau of Indian Standards.
- (49) Rango, T., Kravchenko, J., & Atlaw, B. (2012). Fluoride mobility in East African Rift. *Science of the Total Environment*, 435–436, 130–141. doi: 10.1016/j.scitotenv.2012.07.019
- (50) Chae, G.T., Yun, S.T., Kwon, M.J., Mayer, B., &

- Kim, K.H. (2007). Fluorine geochemistry in South Korea. *Applied Geochemistry*, 22(4), 683–708. doi: 10.1016/j.apgeochem.2006.11.013
- (50) Dean, H.T. (1942). The Investigation of Physiological Effects by the Epidemiological Method. In *Fluorine and Dental Health*.