

Article

Groundwater Fluoride Prediction for Sustainable Water Management: A Comparative Evaluation of Machine Learning Approaches Enhanced by Satellite Embeddings

Yunbo Wei ¹, Rongfu Zhong ² and Yun Yang ^{1,*} ¹ School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China² Zhejiang Environmental Technology Co., Ltd., Hangzhou 311000, China

* Correspondence: yy_hhu@hhu.edu.cn

Abstract

Groundwater fluoride contamination poses a significant threat to sustainable water resources and public health, yet conventional water quality analysis is both time-consuming and costly, making large-scale, sustainable monitoring challenging. Machine learning methods offer a promising, cost-effective, and sustainable alternative for assessing the spatial distribution of fluoride. This study aimed to develop and compare the performance of Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) models for predicting groundwater fluoride contamination in the Datong Basin with the help of satellite embeddings from the AlphaEarth Foundation. Data from 391 groundwater sampling points were utilized, with the dataset partitioned into training (80%) and testing (20%) sets. The ANOVA F-value of each feature was calculated for feature selection, identifying surface elevation, pollution, population, evaporation, vertical distance to the rivers, distance to the Sanggan river, and nine extra bands from the satellite embeddings as the most relevant input variables. Model performance was evaluated using the confusion matrix and the area under the receiver operating characteristic curve (ROC-AUC). The results showed that the SVM model demonstrated the highest ROC-AUC (0.82), outperforming the RF (0.80) and MLP (0.77) models. The introduction of satellite embeddings improved the performance of all three models significantly, with the prediction errors decreasing by 13.8% to 23.3%. The SVM model enhanced by satellite embeddings proved to be a robust and reliable tool for predicting groundwater fluoride contamination, highlighting its potential for use in sustainable groundwater management.

Keywords: Random Forest; Artificial Neural Network; Support Vector Machine; AlphaEarth

Academic Editor: Hossein Bonakdari

Received: 15 August 2025

Revised: 18 September 2025

Accepted: 18 September 2025

Published: 22 September 2025

Citation: Wei, Y.; Zhong, R.; Yang, Y. Groundwater Fluoride Prediction for Sustainable Water Management: A Comparative Evaluation of Machine Learning Approaches Enhanced by Satellite Embeddings. *Sustainability* **2025**, *17*, 8505. <https://doi.org/10.3390/su17188505>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Groundwater provides drinking water for an estimated 1.5 to 3 billion people and accounts for approximately 50% of all domestic water withdrawals worldwide [1–4]. Ensuring the safety and long-term availability of this resource is a cornerstone of sustainable development and global public health [5–7]. The utility of this resource is frequently limited by its chemical composition, and fluoride is one such chemical constituent of concern due to its effects on human health [8–10]. While concentrations ranging from 0.5 mg/L to 1.0 mg/L are considered beneficial for dental health, long-term consumption of water exceeding the World Health Organization (WHO) guideline of 1.5 mg/L [11] (or the Chinese standard of 1.0 mg/L [12]) is associated with adverse health outcomes. These include dental and

skeletal fluorosis, as well as other systemic disorders. High-fluoride groundwater provinces have been identified on every continent, from the East African Rift Valley to the plains of India and the basins of China [13–18], making the large-scale mapping of fluoride risk a critical priority for public health and sustainable water management [13,19].

This global challenge is acutely manifested across the arid and semi-arid basins of North China, where a convergence of geogenic factors and intense water resource exploitation creates widespread fluoride contamination hotspots [19–23]. Among these, the Datong Basin has emerged as a focal point for hydrogeochemical research due to the severity and complexity of its fluoride problem [24]. Extensive studies in the basin have established a clear understanding of the primary drivers: groundwater is naturally enriched with fluoride from the dissolution of fluorine-bearing minerals [25], a process significantly amplified by the region's high evaporation rates [26] and further exacerbated by anthropogenic pressures [27]. In particular, decades of groundwater over-extraction for agriculture have been shown to alter local flow regimes and geochemical balances, accelerating the release and accumulation of fluoride in the aquifer system [28]. Despite this progress in understanding the causal mechanisms, the resulting spatial distribution of high-fluoride groundwater is highly heterogeneous, making large-scale risk assessment a persistent challenge for local water authorities.

The concentration of fluoride in groundwater is controlled by a combination of natural and anthropogenic factors [29–31]. Geogenic sources include the dissolution of fluoride-bearing minerals such as fluorite (CaF_2), apatite, and biotite [19,32], with its mobility governed by local hydrogeochemistry (e.g., pH, alkalinity, ion exchange) [27,33]. In arid and semi-arid climates, evapotranspiration can further increase solute concentrations [33,34]. Anthropogenic inputs from sources like phosphate fertilizers and irrigation return flows may also contribute to fluoride loading in groundwater [31]. The complexity of these interacting factors makes the spatial prediction of fluoride concentrations challenging [32,35]. Process-based numerical models exist for simulating contaminant transport. However, these models have significant limitations for regional-scale applications. They typically require extensive datasets for calibration and validation, are computationally intensive, and involve complex parameterization. These challenges can limit their practical use in creating sustainable and scalable monitoring programs [36,37].

As an alternative to process-based modeling, data-driven machine learning (ML) methods offer an efficient approach for predicting groundwater contamination [38–40]. These algorithms can identify complex, non-linear relationships between a set of predictor variables and a target variable without requiring explicit definition of the underlying physical processes [41,42]. Several ML algorithms have already been applied in hydrogeology. Artificial Neural Networks (ANNs) can model complex relationships in large datasets and have been applied to predict nitrate leaching, hydrological variables, and groundwater quality [43–45]. The Random Forest (RF) algorithm, an ensemble method, is effective for handling high-dimensional data and is resistant to overfitting during the task of water quality classification and nitrate/arsenic concentration prediction [46–48]. Support Vector Machine (SVM) is well-suited for classification tasks, such as differentiating between contaminated and uncontaminated sites, and has been employed to predict nitrate concentrations in groundwater, as well as other contaminants such as sodium and arsenic [49–51].

Despite their success, a critical research gap persists: the majority of these studies rely heavily on hydrochemical input features (e.g., pH, EC, TDS, major ions) for prediction [38,52]. This creates a paradoxical situation where, to predict contamination at a location, one must first collect and analyze a water sample from that location, largely defeating the purpose of a truly predictive, scalable monitoring tool [53]. Thus, alternative

approaches are needed that utilize input variables (such as those from remote sensing) which are readily and inexpensively available over large regions.

This study hypothesizes that this gap can be bridged by leveraging advanced, multi-modal satellite embeddings, such as those from the AlphaEarth Foundation [54], that can serve as powerful proxies for the complex surface and near-surface conditions that control subsurface hydrogeochemistry. These embeddings synthesize vast amounts of data (from optical and radar imagery to elevation and climate models) into a rich, geographically-aware feature set that may implicitly capture signatures of geology, soil type, land use, and human activity [55]. While their performance has been proven in surface-level tasks, their potential to infer subsurface water quality remains a critical and unexplored frontier. This leads to the following primary research questions:

1. Can ML models, trained exclusively on publicly available geospatial data and novel satellite embeddings, accurately predict groundwater fluoride risk without relying on traditional hydrochemical inputs?
2. How do the performances of RF, SVM, and ANN models compare in this new predictive framework, and what is the quantifiable benefit of incorporating satellite embeddings?
3. What are the dominant environmental and anthropogenic drivers of fluoride distribution in the Datong Basin, as revealed by the most effective predictive model?

Therefore, the objective of this study is to develop and evaluate a practical and scalable framework for predicting groundwater fluoride contamination in the Datong Basin. By conducting a comparative evaluation of RF, SVM, and ANN models enhanced with satellite embeddings, this study aims not only to identify the most accurate predictive tool, but also to validate a new, more sustainable approach for regional-scale groundwater risk assessment, thus helping to implement sustainable groundwater management, protect community health, and ensure the long-term security of this vital resource.

2. Study Area Setting

The Datong Basin is a slender graben basin with a northeast–southwest trend situated in the middle-western part of North China (39°54′–40°44′ N, 112°06′–114°33′ E). The basin covers an area of approximately 6000 km² and constitutes a well-defined geomorphological unit, enclosed by the Heng, Cailiang, Liulian, Hongtao, and Guancen mountain ranges [15]. The terrain within the basin is relatively flat, with altitudes ranging from 800 to 1200 m, and is primarily composed of alluvial and fan-deltaic plains. The basin is filled with Quaternary sediments that can reach several hundred meters in thickness [56]. The presence of distinct fault lines and dormant volcanoes in the northern and northwestern sectors indicates that the basin's formation is closely linked to tectonic activity [57].

The region is characterized by a temperate continental monsoon climate, defined by arid conditions with cold, long winters and hot summers. According to meteorological records from the past two decades, the mean annual temperature is 6.8 °C to 8.8 °C [58]. The mean annual precipitation is low, ranging from 300 mm to 400 mm, with over 80% concentrated in the summer months (June to September). In contrast, the mean annual potential evaporation is substantially higher, ranging from 1500 mm to 2000 mm. This significant water deficit underscores the region's aridity and its reliance on groundwater resources [59].

The regional groundwater flow system is generally directed from the piedmont recharge zones at the basin margins towards the center, and from the northwest to the southeast. The primary surface water body is the ephemeral Sanggan River [60]. Agricultural practices include two large-scale irrigation periods annually in March and September.

Furthermore, soil salinization is a significant environmental issue, affecting an estimated 25% to 30% of the basin's land area [61].

3. Materials and Methods

3.1. Predictor Variables and Data Preparation

3.1.1. Target Variable: Groundwater Fluoride Concentration

The dependent variable for this study was groundwater fluoride concentration. The dataset consisted of analytical results from 391 groundwater samples collected across the Datong Basin (Figure 1) in the time period of 2022–2023. Fluoride concentrations were determined in the laboratory using an ion chromatograph (Dionex ICS-600, Thermo Fisher Scientific, MA, USA), a technique widely recognized for its accuracy and reliability in analyzing ionic species in aqueous samples [62]. The testing method strictly followed the Chinese National Environmental Protection Standard for water quality analysis (HJ 84-2016). According to this standard, samples were first filtered through 0.45 μm membrane filters and stored at 4 $^{\circ}\text{C}$ in polyethylene bottles prior to analysis. The analytical process involved separating anions with an anion exchange column (IonPac AS22 with 4.5 mM Na_2CO_3 and 1.4 mM NaHCO_3 eluent) and quantifying them using a suppressed conductivity detector. The method detection limit (MDL) for fluoride was 0.02 mg/L. Rigorous quality assurance and quality control (QA/QC) procedures were implemented, including the analysis of duplicate samples, blanks, and standard reference materials. The relative percent difference (RPD) for duplicate samples was consistently below 5%, and the recovery of spiked samples ranged from 95% to 105%, ensuring the accuracy and reliability of the dataset. The continuous fluoride concentration data was transformed into a binary classification problem by categorizing samples as either 'low-risk' (≤ 1.0 mg/L, class 0) or 'high-risk' (> 1.0 mg/L, class 1). This was a deliberate methodological choice designed to align the model's output directly with the critical regulatory threshold for safe drinking water. From a practical standpoint, this framework mirrors the binary decision-making process of water managers, whose primary concern is identifying sources that exceed permissible limits. From a technical standpoint, this simplification focuses the learning algorithm on distinguishing safe from unsafe water, while making the model more robust to minor fluctuations and noise. This approach ultimately yields a more stable, accurate, and actionable tool for groundwater quality assessment. Subsequently, for model training and validation, the dataset was subjected to a standard 80/20 random split, creating a 80% training subset and a hold-out 20% testing subset.

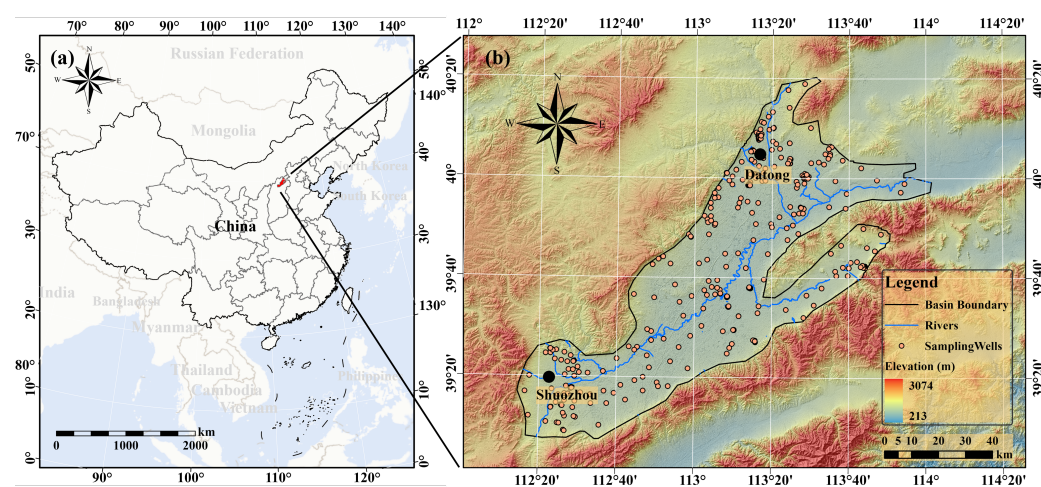


Figure 1. Localization of the study area: (a) localization of China and Datong Basin; (b) Digital Elevation Model of hydrographic boundaries of Datong Basin, and location of sampling wells.

3.1.2. Predictor Variable Datasets

A comprehensive suite of predictor variables was compiled, categorized as meteorological, topographic, hydrological, soil, anthropogenic, and satellite embeddings.

Meteorological Variables: Gridded datasets for annual mean precipitation, annual mean evaporation, and annual mean air temperature were obtained from the Center for Resource and Environmental Sciences and Data, Chinese Academy of Sciences (www.resdc.cn), at an initial resolution of 1 km.

Topographical Variables: A Digital Elevation Model (DEM) with a 30 m resolution was acquired from the Geospatial Data Cloud (www.gscloud.cn). This DEM was used to derive the ground surface elevation and ground surface slope. Proximity variables, including the distance to the Sanggan River, the vertical and horizontal distance to the nearest river, and the distance to urban and arable land, were calculated in ArcGIS Pro 2.8 using river network and land use maps as reference layers.

Hydrological Variables: Key hydrogeological parameters were compiled from field data and existing reports. Hydraulic conductivity was derived from pumping test data. The infiltration rate was estimated based on soil and land use characteristics. Groundwater table elevation and depth were interpolated from measurements at 391 observation wells. Information on the soil type of the vadose zone, the soil type of the aquifer, and aquifer thickness was obtained from the lithological descriptions of 130 well logs.

Soil Variables: Eight soil properties were obtained from the ISRIC—World Soil Information database (SoilGrids v2.0, <https://www.isric.org/>) at a 250 m spatial resolution. The variables included the following: bulk density, coarse fragment content, sand content, silt content, cation exchange capacity (at pH 7), soil organic carbon, soil organic carbon stock, and nitrogen content.

Anthropogenic Variables: Data representing human impact were also included. Population density was sourced from the Baidu Huiyan dataset (<https://huiyan.baidu.com/>). The pollution overload index was calculated based on the distribution of potential industrial and agricultural contamination sources.

Satellite Embeddings: In addition to these traditional variables, 64 satellite embedding features from the AlphaEarth Foundation (<https://developers.google.com/earth-engine/datasets> (accessed on 10 September 2025)) were incorporated in the prediction models. These are not direct physical measurements; instead, they are 64-dimensional feature vectors produced by a self-supervised deep learning model. The model synthesizes amounts of multi-modal data—including optical (Sentinel-2, Landsat), radar (Sentinel-1), elevation, and climate data—into a compact and semantically rich representation of the Earth’s surface. This approach is designed to capture complex, non-linear relationships across diverse datasets, offering a more holistic characterization of surface conditions relevant to groundwater processes [54,55].

3.1.3. Determination of the Most Relevant Input Variables

Effective feature selection is crucial for optimizing model performance by eliminating irrelevant variables [63]. To identify the most influential predictors for this classification task, an Analysis of Variance (ANOVA) F-test was employed. This statistical test assesses the discriminatory power of each feature by calculating the ratio of variance between the high- and low-fluoride classes to the variance within each class. A higher F-value signifies greater class separation, indicating a more significant feature. The F-statistic formula is defined as follows:

$$F = \frac{\text{Variance between the high- and low-fluoride classes}}{\text{Variance within each class}} \quad (1)$$

Based on this analysis, the top 15 features with the highest F-values were selected for model training. When satellite embeddings were used for prediction, these features included six environmental and geospatial variables (surface elevation, pollution, population, evaporation, vertical distance to rivers, and distance to the Sanggan river) and nine extra bands from the satellite embeddings. While when satellite embeddings were not used, the models utilized a set consisting exclusively of the top 15 environmental and geospatial variables (Table 1).

Table 1. Selection of most relevant inputs with and without satellite embeddings.

Embedding-Enhanced Models		Conventional Models	
Variables	ANOVA F-Value	Variables	ANOVA F-Value
surface elevation	29.1081	surface elevation	29.1081
pollution	16.1294	pollution	16.1294
population	13.1675	population	13.1675
evaporation	12.9844	evaporation	12.9844
vertical distance to rivers	12.9710	vertical distance to rivers	12.9710
distance to Sanggan river	12.0708	distance to Sanggan river	12.0708
satellite embedding A38	27.2939	vadose zone soil	9.9553
satellite embedding A33	25.7531	precipitation	9.4695
satellite embedding A34	21.4152	hydraulic conductivity	6.9283
satellite embedding A41	20.5796	distance to urban area	4.8268
satellite embedding A25	18.6943	groundwater table elevation	4.3817
satellite embedding A23	17.9126	horizontal distance to rivers	3.4892
satellite embedding A14	13.4771	soil organic carbon	3.1402
satellite embedding A20	12.2590	bulk density	2.4634
satellite embedding A11	11.5991	air temperature	2.0761

3.2. Model Development

This study employed a comparative approach, evaluating three benchmark machine learning algorithms: ANN, RF, and SVM. The selection of these specific models was deliberate and methodological, intended to represent three distinct and highly successful families of machine learning paradigms [64]. The ANN represents the foundational neural network approach, capable of capturing complex, hierarchical non-linear patterns. RF represents ensemble learning (specifically, bagging), which is known for its robustness and resistance to overfitting. Finally, SVM represents kernel-based methods, which excel at finding optimal decision boundaries in high-dimensional spaces. By comparing these diverse approaches, the best-suited learning philosophy for this specific hydrogeological prediction task can be determined. While other powerful algorithms like XGBoost (a boosting-based ensemble) and deep Convolutional Neural Networks (CNNs) exist, they were not selected for this primary comparison. XGBoost, while powerful, belongs to the same family of tree-based ensembles as RF. Compared with XGBoost, RF is generally less sensitive to hyperparameter tuning and more robust against overfitting compared to boosting models [64,65], making it an ideal candidate for establishing a stable and reliable baseline in a comparative study like this one. CNNs are primarily designed for grid-structured data (e.g., images) and are less directly applicable to the point-based sampling dataset without significant preprocessing that would introduce its own set of assumptions. Therefore, the chosen trio provides a robust and well-established baseline for evaluating the utility of satellite embeddings in this context. The machine learning models were built in Python 3.10 using the Scikit-Learn library.

3.2.1. ANN

A Multilayer Perceptron (MLP), a foundational class of feedforward ANNs [66], was implemented as the first model. The model was configured with two hidden layers, with the number of neurons in the input and output layers determined by the number of predictor variables and the binary classification target, respectively. To optimize the network's weights during training, the Adam optimizer was selected. This choice was justified by its proven computational efficiency and widespread adoption for its effectiveness in a variety of deep learning applications [38,67].

3.2.2. RF

The RF model, an ensemble learning algorithm built from a multitude of decision trees [68], was utilized as the second model. To maximize its predictive power, a systematic hyperparameter tuning process was conducted, which is a critical step for optimizing model performance [65]. A grid search was performed to identify the optimal combination of key hyperparameters based on the highest ROC-AUC achieved during cross-validation. The final optimized architecture consisted of 55 trees (*n_estimators*), a maximum tree depth of eight (*max_depth*), and two random features considered at each split (*max_features*) to ensure tree diversity and reduce variance.

3.2.3. SVM

The SVM classifier, a powerful method for finding an optimal separating hyperplane [69], was implemented as the third model. The Radial Basis Function (RBF) Kernel was specifically chosen for this task. The RBF Kernel is highly effective at handling complex, non-linear relationships between predictors and is widely recommended as a robust default choice for classification problems where the data structure is not linearly separable [70]. This kernel-based approach allows the model to operate in a high-dimensional feature space, enabling it to identify a non-linear decision boundary between the high- and low-fluoride classes.

3.2.4. Model Validation

The performance of each trained model was assessed on the independent testing set using several standard metrics derived from a confusion matrix. The confusion matrix tabulated the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). The primary evaluation metrics were Accuracy, Precision, Recall, Specificity, F1 score, and the area under the receiver operating characteristic curve (ROC-AUC) [38].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$F1\ score = \frac{2\ TP}{2\ TP + FP + FN} \quad (6)$$

where true positive (*TP*) indicates that the high-fluoride sample is correctly identified; true negative (*TN*) indicates that the low-fluoride sample is correctly identified; false positive (*FP*) represents the misclassification of a low-fluoride sample as high; and false negative (*FN*) represented the misclassification of a high-fluoride sample as low.

The ROC-AUC is particularly well-suited for this type of environmental application, as it provides a threshold-independent measure of a model's ability to distinguish between classes and is robust to class imbalance, which can be present in contamination datasets [71]. Apart from ROC-AUC, the combination of the Accuracy, Precision, Recall, Specificity, and F1 score metrics provides a holistic view of model performance. These metrics capture not only the overall correctness (Accuracy), but also the trade-offs between different types of classification errors [14,72].

3.2.5. Feature Importance Analysis

To interpret the model and identify the key drivers of fluoride contamination, a feature importance analysis was conducted. The Random Forest (RF) model was selected for this task due to its inherent ability to provide feature importance metrics, offering greater interpretability compared to the SVM and ANN models [38]. The importance of each predictor variable was quantified using the Mean Decrease in Impurity (MDI), also known as Gini importance [73].

The MDI method evaluates the importance of a feature by measuring how much it contributes to reducing uncertainty within the decision trees of the RF ensemble. This uncertainty is quantified by the Gini impurity. For a given node m in a tree, the Gini impurity is calculated as follows:

$$G(m) = \sum_{k=1}^K p_{mk}(1 - p_{mk}) = 1 - \sum_{k=1}^K p_{mk}^2 \quad (7)$$

where K is the number of classes (in this study, $K = 2$), and p_{mk} is the proportion of samples belonging to class k at node m . A Gini impurity of 0 indicates that the node is pure (all samples belong to a single class).

When a feature is used to split a parent node m into two child nodes (left and right), the quality of the split is measured by the decrease in Gini impurity, calculated as follows:

$$\Delta G(m) = G_{parent} - \left(\frac{N_{left}}{N_{parent}} G_{left} + \frac{N_{right}}{N_{parent}} G_{right} \right) \quad (8)$$

where N_{parent} , N_{left} , and N_{right} are the number of samples in the parent, left child, and right child nodes, respectively.

The importance of a single feature in one tree is the sum of the Gini impurity decreases for all nodes where that feature was used for splitting. The final MDI score for the feature is the average of these importance values across all the trees in the forest. A higher MDI value signifies that the feature is more effective at partitioning the data into the defined classes and is therefore considered more influential in the model's predictions.

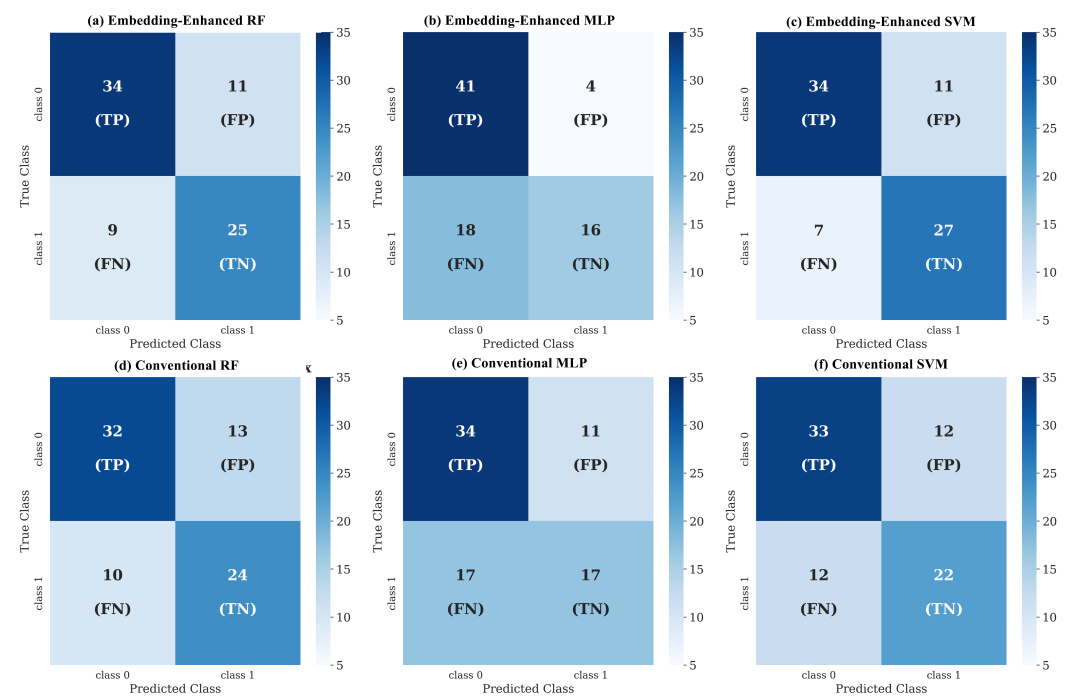
4. Results

4.1. Model Evaluation and Comparison

Following the training phase, the performance of all six model configurations (RF, ANN, and SVM, with and without satellite embeddings) was evaluated on the independent test set. The resulting metrics (Table 2) and confusion matrices (Figure 2) demonstrate a clear performance hierarchy. The embedding-enhanced models consistently outperformed their conventional counterparts, with prediction errors decreasing by 13.8% to 23.3%.

Table 2. Performance metrics for the RF, ANN, and SVM models on the testing dataset.

No.	Embedding-Enhanced Models			Conventional Models		
	RF	ANN	SVM	RF	ANN	SVM
Accuracy	0.75	0.72	0.77	0.71	0.65	0.70
Precision	0.69	0.80	0.71	0.65	0.61	0.65
Recall	0.74	0.47	0.79	0.71	0.50	0.65
Specificity	0.76	0.91	0.76	0.71	0.76	0.73
F1 score	0.71	0.59	0.75	0.68	0.55	0.65
ROC-AUC	0.80	0.77	0.82	0.78	0.74	0.77

**Figure 2.** Confusion matrix: with satellite embeddings—(a) RF model, (b) ANN model, and (c) SVM model; without satellite embeddings—(d) RF model, (e) ANN model, and (f) SVM model. The values on the top left, bottom right, top right, and bottom left represent the number of TPs, TNs, FPs, and FNs, respectively.

Among the models, the embedding-enhanced SVM model emerged as the top performer, achieving an accuracy of 0.77, a recall of 0.79, and an ROC-AUC of 0.82. The embedding-enhanced RF model also demonstrated strong predictive capability with an accuracy of 0.75 and an ROC-AUC of 0.80. The ANN model yielded the lowest metrics of the three, with an accuracy of 0.72 and an ROC-AUC of 0.77. Its performance was primarily hindered by a low recall of 0.47.

4.2. Variables Influencing Fluoride Distribution

A feature importance analysis was conducted using the trained RF model to quantify the relative influence of each predictor variable, as detailed in Section 3.2.5. The analysis, summarized in Figure 3, identified the most significant drivers of fluoride distribution. Among meteorological factors, annual mean evaporation (EVP) emerged as a highly influential predictor (MDI = 0.1008). Topography also played a critical role, with ground surface elevation (DEM) being the most influential predictor (MDI = 0.0983), followed by vertical distance to rivers (DrivfdirV) and distance to the Sanggan River (Dsanggan). Furthermore, population also emerged as a significant anthropogenic predictor of groundwater fluoride.

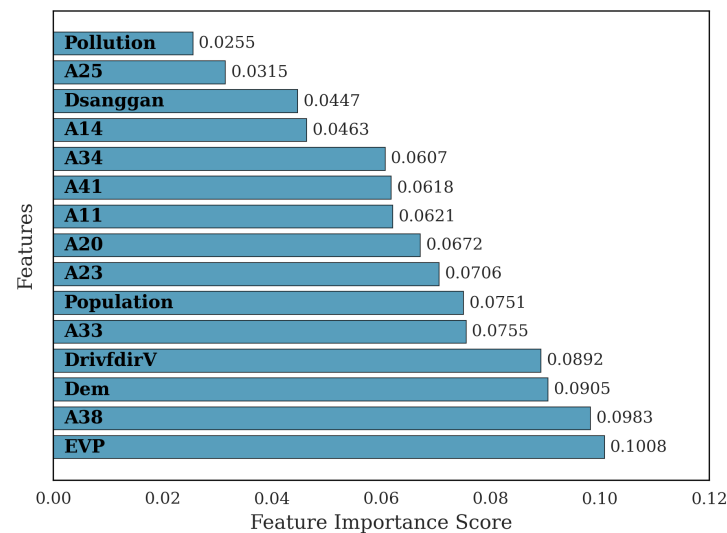


Figure 3. Feature importance, evaluated using MDI.

4.3. Predicted Spatial Distribution of Groundwater Fluoride

Leveraging the superior performance of the embedding-enhanced SVM model, a predictive risk map for groundwater fluoride across the Datong Basin was generated (Figure 4). The map highlights a pronounced spatial heterogeneity in fluoride distribution, identifying the central and northern parts of the plain as the primary high-risk zones ($F^- > 1.0$ mg/L).

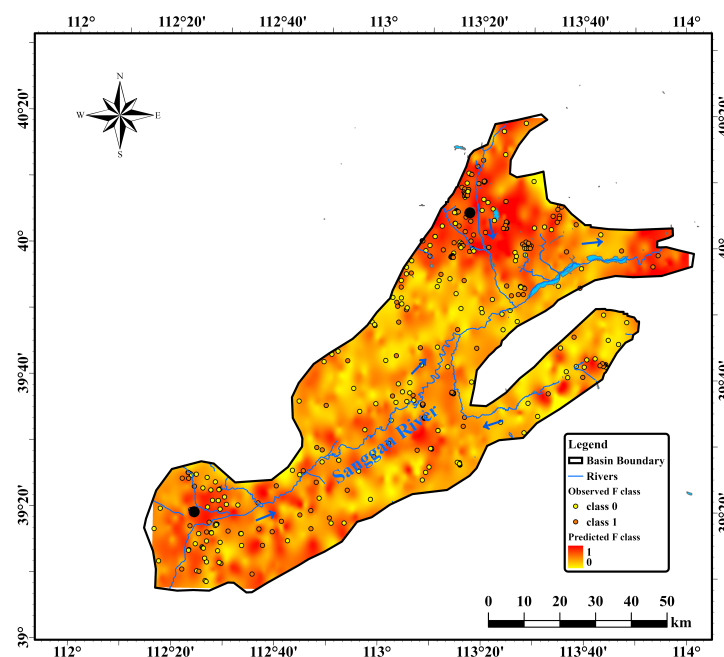


Figure 4. Spatial distribution of groundwater fluoride, predicted by embedding-enhanced SVM model.

5. Discussion

5.1. Interpretation of Model Performance

The superior performance of the SVM is attributable to its fundamental design. SVMs excel at finding the optimal hyperplane that best separates classes in a high-dimensional

feature space [69]. By using a Radial Basis Function (RBF) Kernel, the model effectively captures complex and non-linear relationships between predictors, which is essential for modeling the intricate hydrogeochemical systems of the Datong Basin [49]. Its effectiveness in this high-dimensional context highlights its robustness against overfitting compared to other methods [51,74]. The RF model's strong performance is also consistent with its established suitability for complex environmental datasets [75,76]. As an ensemble method, RF inherently minimizes overfitting and captures non-linear predictor interactions, making it a reliable tool for this type of classification task [14,77].

In contrast, the ANN model showed the weakest performance. While ANNs are powerful, their performance can be highly sensitive to network architecture, hyperparameter tuning, and the size of the training dataset [49,78]. In this instance, the model may have struggled to generalize as effectively as the SVM or RF, potentially due to the dataset's complexity relative to its size, a known challenge for neural networks.

In summary, this study demonstrates that while all three algorithms have utility, the SVM model, when enhanced with satellite embeddings, provides the most accurate and reliable tool for this classification task. Two factors contribute to the model's predictive accuracy: first, the algorithm's inherent ability to capture complex, non-linear relationships; second, the predictor set formed by integrating traditional variables with satellite embeddings.

5.2. Hydrogeological Significance of Key Drivers

The feature importance analysis provides critical insights into the physical processes governing fluoride distribution. The model's identification of key predictors aligns remarkably well with established hydrogeological knowledge of the Datong Basin, lending strong credibility to its findings.

The high importance score of EVP is strongly supported by the literature. The Datong Basin is characterized by an arid-to-semi-arid climate with high potential evaporation (often exceeding 2000 mm/year) and low precipitation [25,26]. This significant water deficit makes evapoconcentration a dominant hydrogeochemical process. Numerous studies have recognized that this strong evaporation is a key factor in the enrichment of various solutes, including fluoride, in the basin's shallow groundwater [26,27,79]. Therefore, the model's identification of EVP as an important predictor is consistent with the fundamental hydrogeological processes of the region.

The importance of topographic variables, particularly DEM and distances to rivers, is also deeply rooted in the basin's hydrogeology. The established conceptual model for the Datong Basin describes a flow system where groundwater is recharged in the higher-elevation piedmont areas and flows towards the central, lower-elevation plains, which act as discharge zones [58,60]. Research indicates that fluoride, after being leached from fluorine-bearing minerals in sediments and strata under alkaline conditions [12], is transported by this groundwater flow and accumulates in these low-elevation discharge zones [24]. Here, the water table is closer to the surface, allowing the accumulated fluoride to be further concentrated by the intense evaporation discussed above, leading to significantly elevated levels [26,27].

Finally, the model's identification of population as a significant predictor is a powerful reflection of anthropogenic pressures. Population density serves as a direct proxy for water demand for both domestic and agricultural purposes. Previous studies have extensively documented the consequences of these demands in the Datong Basin, including long-term over-extraction of groundwater, the formation of significant cones of depression in the water table, and a continuous decline in groundwater levels since the 1980s [24,57,60]. This intensive pumping has been identified as a dominant mechanism that induces the release

of fluoride ions from surrounding clayey soils into the aquifer, particularly in the northern parts of the basin [24]. Thus, the correlation found by the model between population and fluoride risk is strongly supported by documented anthropogenic impacts on the local aquifer system.

5.3. Implications of the Predicted Spatial Distribution

The predictive risk map generated by the embedding-enhanced SVM model (Figure 4) provides a spatially explicit confirmation of the hydrogeological drivers discussed above. The model identifies the central and northern parts of the plain as the primary high-risk zones, a pattern that aligns remarkably well with spatial distributions observed in other studies [26,58]. This concentration of high-fluoride groundwater along river systems and within the basin's low-lying northern discharge zone confirms that these are the areas where the effects of natural solute transport, intense evaporation, and anthropogenic groundwater over-extraction are most acute [24,28]. This strong correspondence between the data-driven model's output and established process-based understanding further validates the model's predictive capabilities and its utility for sustainable water management.

6. Conclusions

This study successfully developed and validated a machine learning framework for predicting high-fluoride groundwater risk in the Datong Basin by integrating conventional geospatial data with novel satellite embeddings. A comparative analysis of RF, ANN, and SVM models was carried out. Among the models tested, the SVM model, enhanced with a Radial Basis Function (RBF) Kernel, emerged as the most robust and accurate predictor, achieving the highest accuracy (0.77) and ROC-AUC (0.82) on the independent test set. Furthermore, the inclusion of satellite embeddings from the AlphaEarth Foundation improved the predictive power across all models, reducing prediction errors by 13.8% to 23.3%. Finally, the feature importance analysis confirmed that the model captures physically meaningful processes. Evaporation, surface elevation, and population density were identified as the dominant drivers of fluoride enrichment, aligning with the established hydrogeological understanding that fluoride accumulation in the Datong Basin is controlled by a combination of evapoconcentration in discharge zones and anthropogenic pressures from groundwater extraction.

Despite these promising results, several limitations should be acknowledged. The primary limitation is the inherent 'black box' nature of the satellite embeddings. Although this study demonstrated their validity for prediction, a gap remains between this effectiveness and a clear physical explanation. Secondly, the model's performance has been validated only within the specific hydrogeological context of the Datong Basin, and its transferability to regions with different geological settings or climatic conditions remains untested. Finally, the binary classification framework (high-risk vs. low-risk), while practical for regulatory purposes, simplifies the problem and does not predict the continuous spectrum of fluoride concentrations.

Future research should be directed toward addressing these limitations and expanding upon the current framework. Firstly, employing explainable machine learning techniques, could help to 'open the black box' and reveal which specific surface features captured by the embeddings are most influential. Secondly, future work should focus on testing and adapting the model in diverse geographical and geological settings to build more universally applicable prediction tools. Finally, transitioning from a classification- to a regression-based approach would enable the prediction of actual fluoride concentration values, providing a more granular and informative risk map for water managers and public health officials.

Author Contributions: Conceptualization, Y.W., R.Z. and Y.Y.; Formal analysis, Y.W., R.Z. and Y.Y.; Funding acquisition, Y.W. and Y.Y.; Investigation, Y.W., R.Z. and Y.Y.; Methodology, Y.W., R.Z. and Y.Y.; Project administration, Y.Y.; Supervision, Y.Y.; Validation, Y.W., R.Z. and Y.Y.; Visualization, Y.W. and R.Z.; Writing—original draft, Y.W. and R.Z.; Writing—review & editing, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Key Research and Development Program of China [grant number 2023YFC3209700], the National Natural Science Foundation of China [grant number 42102282], and the Natural Science Foundation of Jiangsu Province [grant number BK20210378].

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We express our deepest gratitude to the editors and anonymous reviewers for their careful work and insightful comments that helped to improve this paper.

Conflicts of Interest: Author Rongfu Zhong was employed by the company Zhejiang Environmental Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Margat, J.; Van der Gun, J. *Groundwater Around the World: A Geographic Synopsis*; CRC Press: Boca Raton, USA, 2013; pp. 148–149.
2. Beyene, G.; Abera, D.; Fufa, F. Evaluation of the suitability of groundwater for drinking and irrigation purposes in Jimma Zone of Oromia, Ethiopia. *Groundw. Sustain. Dev.* **2019**, *9*, 100216. [[CrossRef](#)]
3. United Nations Educational, Scientific and Cultural Organization (UNESCO). *The United Nations World Water Development Report 2022: Groundwater: Making the Invisible Visible*; Technical report; UNESCO: Paris, France, 2022.
4. Nazari, S.; Reinecke, R.; Moosdorf, N. Global estimates of groundwater withdrawal trends and uncertainties. *Environ. Res. Lett.* **2025**, *20*, 094043. [[CrossRef](#)]
5. Gleeson, T.; Alley, W.M.; Allen, D.M.; Sophocleous, M.A.; Zhou, Y.; Taniguchi, M.; VanderSteen, J. Towards sustainable groundwater use: Setting long-term goals, backcasting, and managing adaptively. *Groundwater* **2012**, *50*, 19–26.
6. Shaikh, M.; Birajdar, F. Groundwater and ecosystems: Understanding the critical interplay for sustainability and conservation. *EPRA Int. J. Multidiscip. Res.* **2024**, *10*, 181–186.
7. World Health Organization. *World Health Statistics 2025: Monitoring Health for the SDGs, Sustainable Development Goals*; World Health Organization: Geneva, Switzerland, 2025; pp. 38–39.
8. Kimambo, V.; Bhattacharya, P.; Mtalo, F.; Mtamba, J.; Ahmad, A. Fluoride occurrence in groundwater systems at global scale and status of defluoridation-state of the art. *Groundw. Sustain. Dev.* **2019**, *9*, 100223. [[CrossRef](#)]
9. Fawell, J.K.; Bailey, K. *Fluoride in Drinking-Water*; World Health Organization: Geneva, Switzerland, 2006; pp. 2–3.
10. Ayoob, S.; Gupta, A.K. Fluoride in drinking water: A review on the status and stress effects. *Crit. Rev. Environ. Sci. Technol.* **2006**, *36*, 433–487. [[CrossRef](#)]
11. World Health Organization. *Guidelines for Drinking-Water Quality*, 4th ed.; World Health Organization: Geneva, Switzerland, 2011; pp. 41–42.
12. Su, C.; Wang, Y.; Xie, X.; Li, J. Aqueous geochemistry of high-fluoride groundwater in Datong Basin, Northern China. *J. Geochem. Explor.* **2013**, *135*, 79–92. [[CrossRef](#)]
13. Podgorski, J.E.; Labhasetwar, P.; Saha, D.; Berg, M. Prediction modeling and mapping of groundwater fluoride contamination throughout India. *Environ. Sci. Technol.* **2018**, *52*, 9889–9898. [[CrossRef](#)]
14. Nafouanti, M.B.; Li, J.; Mustapha, N.A.; Uwamungu, P.; Al-Alimi, D. Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: Comparison of random forest, logistic regression and artificial neural network. *Appl. Geochem.* **2021**, *132*, 105054. [[CrossRef](#)]
15. Nafouanti, M.B.; Li, J.; Nyakilla, E.E.; Mwakipunda, G.C.; Mulashani, A. A novel hybrid random forest linear model approach for forecasting groundwater fluoride contamination. *Environ. Sci. Pollut. Res.* **2023**, *30*, 50661–50674. [[CrossRef](#)]
16. Rafique, T.; Naseem, S.; Bhanger, M.I.; Usmani, T.H. Fluoride ion contamination in the groundwater of Mithi sub-district, the Thar Desert, Pakistan. *Environ. Geol.* **2008**, *56*, 317–326. [[CrossRef](#)]
17. Tekle-Haimanot, R.; Melaku, Z.; Kloos, H.; Reimann, C.; Fantaye, W.; Zerihun, L.; Bjorvatn, K. The geographic distribution of fluoride in surface and groundwater in Ethiopia with an emphasis on the Rift Valley. *Sci. Total Environ.* **2006**, *367*, 182–190.
18. Borgnino, L.; Garcia, M.; Bia, G.; Stupar, Y.; Le Coustumer, P.; Depetris, P. Mechanisms of fluoride release in sediments of Argentina’s central region. *Sci. Total Environ.* **2013**, *443*, 245–255.

19. Zhang, Z.; Liu, J.; Xiao, Z.; Liu, F.; Wang, Z.; Chen, S.; Zhang, J.; Xia, Y.; Jiang, W.; Ning, H. Spatial distribution, controlling factors, and health risk assessment of groundwater fluoride in the Chahanur Basin, Inner Mongolia, China. *Environ. Earth Sci.* **2025**, *84*, 400. [\[CrossRef\]](#)
20. Li, J.; Wang, Y.; Zhu, C.; Xue, X.; Qian, K.; Xie, X.; Wang, Y. Hydrogeochemical processes controlling the mobilization and enrichment of fluoride in groundwater of the North China Plain. *Sci. Total Environ.* **2020**, *730*, 138877. [\[CrossRef\]](#)
21. Sun, D.; Li, J.; Li, H.; Liu, Q.; Zhao, S.; Huang, Y.; Wu, Q.; Xie, X. Evolution of groundwater salinity and fluoride in the deep confined aquifers of Cangzhou in the North China plain after the South-to-North Water Diversion Project. *Appl. Geochem.* **2022**, *147*, 105485.
22. Cao, W.; Zhang, Z.; Fu, Y.; Zhao, L.; Ren, Y.; Nan, T.; Guo, H. Prediction of arsenic and fluoride in groundwater of the North China Plain using enhanced stacking ensemble learning. *Water Res.* **2024**, *259*, 121848. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Chen, Y.; Hou, J.; Zhou, J.; Yu, J.; Zhang, J.; Zhao, J. Hydrogeochemical Processes and Sustainability Challenges of Arsenic-and Fluoride-Contaminated Groundwater in Arid Regions: Evidence from the Tarim Basin, China. *Sustainability* **2025**, *17*, 7971.
24. Li, L.; Wang, Y.; Wu, Y.; Li, J. Major geochemical controls on fluoride enrichment in groundwater: A case study at Datong Basin, northern China. *J. Earth Sci.* **2013**, *24*, 976–986. [\[CrossRef\]](#)
25. Feng, F.; Jia, Y.; Yang, Y.; Huan, H.; Lian, X.; Xu, X.; Xia, F.; Han, X.; Jiang, Y. Hydrogeochemical and statistical analysis of high fluoride groundwater in northern China. *Environ. Sci. Pollut. Res.* **2020**, *27*, 34840–34861. [\[CrossRef\]](#)
26. Wang, X.; Weerasinghe, R.N.N.; Su, C.; Wang, M.; Jiang, J. Origin and Enrichment Mechanisms of Salinity and Fluoride in Sedimentary Aquifers of Datong Basin, Northern China. *Int. J. Environ. Res. Public Health* **2023**, *20*, 1832. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Su, C.; Wang, Y.; Xie, X.; Zhu, Y. An isotope hydrochemical approach to understand fluoride release into groundwaters of the Datong Basin, Northern China. *Environ. Sci. Processes Impacts* **2015**, *17*, 791–801.
28. Li, J.; Wang, Y.; Xie, X.; Su, C. Hierarchical cluster analysis of arsenic and fluoride enrichments in groundwater from the Datong basin, Northern China. *J. Geochem. Explor.* **2012**, *118*, 77–89. [\[CrossRef\]](#)
29. Sridhar, C.; Thirumurugan, M.; Subramani, T.; Gopinathan, P. Global distribution and sources of uranium and fluoride in groundwater: A comprehensive review. *J. Geochem. Explor.* **2025**, *270*, 107665.
30. Chaudhuri, R.; Sahoo, S.; Debsarkar, A.; Hazra, S. Fluoride Contamination in Groundwater—A Review. In *Geospatial Practices in Natural Resources Management*; Springer International Publishing: Cham, Switzerland, 2024; pp. 331–354.
31. Shaji, E.; Sarath, K.; Santosh, M.; Krishnaprasad, P.; Arya, B.; Babu, M.S. Fluoride contamination in groundwater: A global review of the status, processes, challenges, and remedial measures. *Geosci. Front.* **2024**, *15*, 101734.
32. Jacks, G.; Bhattacharya, P.; Chaudhary, V.; Singh, K. Controls on the genesis of some high-fluoride groundwaters in India. *Appl. Geochem.* **2005**, *20*, 221–228. [\[CrossRef\]](#)
33. Edmunds, W.M.; Smedley, P.L. Fluoride in natural waters. In *Essentials of Medical Geology: Revised Edition*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 311–336.
34. Amini, M.; Mueller, K.; Abbaspour, K.C.; Rosenberg, T.; Afyuni, M.; Möller, K.N.; Sarr, M.; Johnson, C.A. Statistical modeling of global geogenic fluoride contamination in groundwaters. *Environ. Sci. Technol.* **2008**, *42*, 3662–3668. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Chaney, R.L. Food safety issues for mineral and organic fertilizers. *Adv. Agron.* **2012**, *117*, 51–116.
36. Rapantova, N.; Grmela, A.; Vojtek, D.; Halir, J.; Michalek, B. Ground water flow modelling applications in mining hydrogeology. *Mine Water Environ.* **2007**, *26*, 264–270. [\[CrossRef\]](#)
37. Alagha, J.S.; Said, M.A.M.; Mogheir, Y. Modeling of nitrate concentration in groundwater using artificial intelligence approach—a case study of Gaza coastal aquifer. *Environ. Monit. Assess.* **2014**, *186*, 35–45. [\[CrossRef\]](#)
38. Haggerty, R.; Sun, J.; Yu, H.; Li, Y. Application of machine learning in groundwater quality modeling—A comprehensive review. *Water Res.* **2023**, *233*, 119745. [\[CrossRef\]](#)
39. Nadiri, A.A.; Fijani, E.; Tsai, F.T.C.; Asghari Moghaddam, A. Supervised committee machine with artificial intelligence for prediction of fluoride concentration. *J. Hydroinf.* **2013**, *15*, 1474–1490. [\[CrossRef\]](#)
40. Sajedi-Hosseini, F.; Malekian, A.; Choubin, B.; Rahmati, O.; Cipullo, S.; Coulon, F.; Pradhan, B. A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. *Sci. Total Environ.* **2018**, *644*, 954–962. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Bhowmik, T.; Sarkar, S.; Sen, S.; Mukherjee, A. Application of machine learning in delineating groundwater contamination at present times and in climate change scenarios. *Curr. Opin. Environ. Sci. Health* **2024**, *39*, 100554. [\[CrossRef\]](#)
42. Hosseini, F.S.; Choubin, B.; Bagheri-Gavkosh, M.; Karimi, O.; Taromideh, F.; Mako, C. Susceptibility assessment of groundwater nitrate contamination using an ensemble machine learning approach. *Groundwater* **2023**, *61*, 510–516.
43. Baghapour, M.A.; Fadaei Nobandegani, A.; Talebbeydokhti, N.; Bagherzadeh, S.; Nadiri, A.A.; Gharekhani, M.; Chitsazan, N. Optimization of DRASTIC method by artificial neural network, nitrate vulnerability index, and composite DRASTIC models to assess groundwater vulnerability for unconfined aquifer of Shiraz Plain, Iran. *J. Environ. Health Sci. Eng.* **2016**, *14*, 13. [\[CrossRef\]](#)
44. Charulatha, G.; Srinivasalu, S.; Uma Maheswari, O.; Venugopal, T.; Giridharan, L. Evaluation of ground water quality contaminants using linear regression and artificial neural network models. *Arabian J. Geosci.* **2017**, *10*, 128. [\[CrossRef\]](#)

45. Beeralala, A.K.; Gobinath, R.; Shyamala, G.; Manvitha, S. Water quality prediction using statistical tool and machine learning algorithm. In *Waste Management: Concepts, Methodologies, Tools, and Applications*; IGI Global Scientific Publishing: Hershey, PA, USA, 2020; pp. 609–623.
46. Rodriguez-Galiano, V.F.; Luque-Espinar, J.A.; Chica-Olmo, M.; Mendes, M.P. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. *Sci. Total Environ.* **2018**, *624*, 661–672. [[CrossRef](#)]
47. Knoll, L.; Breuer, L.; Bach, M. Nation-wide estimation of groundwater redox conditions and nitrate concentrations through machine learning. *Environ. Res. Lett.* **2020**, *15*, 064004. [[CrossRef](#)]
48. Saghebain, S.M.; Sattari, M.T.; Mirabbasi, R.; Pal, M. Ground water quality classification by decision tree method in Ardebil region, Iran. *Arabian J. Geosci.* **2014**, *7*, 4767–4777. [[CrossRef](#)]
49. Park, Y.; Ligaray, M.; Kim, Y.M.; Kim, J.H.; Cho, K.H.; Sthianopkao, S. Development of enhanced groundwater arsenic prediction model using machine learning approaches in Southeast Asian countries. *Desalin. Water Treat.* **2016**, *57*, 12227–12236.
50. Liu, J.; Gu, J.; Li, H.; Carlson, K.H. Machine learning and transport simulations for groundwater anomaly detection. *J. Comput. Appl. Math.* **2020**, *380*, 112982. [[CrossRef](#)]
51. Isazadeh, M.; Biazar, S.M.; Ashrafzadeh, A. Support vector machines and feed-forward neural networks for spatial modeling of groundwater qualitative parameters. *Environ. Earth Sci.* **2017**, *76*, 610. [[CrossRef](#)]
52. Singh, G.; Mehta, S. Prediction of geogenic source of groundwater fluoride contamination in Indian states: A comparative study of different supervised machine learning algorithms. *J. Water Health* **2024**, *22*, 1387–1408. [[CrossRef](#)]
53. Agrawal, A.; Petersen, M.R. Detecting arsenic contamination using satellite imagery and machine learning. *Toxics* **2021**, *9*, 333. [[CrossRef](#)]
54. Tollefson, J. Google AI model creates maps of Earth ‘at any place and time’. *Nature* **2025**, *644*, 313. [[PubMed](#)]
55. Brown, C.F.; Kazmierski, M.R.; Pasquarella, V.J.; Rucklidge, W.J.; Samsikova, M.; Zhang, C.; Shelhamer, E.; Lahera, E.; Wiles, O.; Ilyushchenko, S.; et al. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv* **2025**, arXiv:2507.22291. [[CrossRef](#)]
56. Zhao, S.; Li, J.; Xue, X.; Sun, D.; Liu, W.; Zhu, C.; Yang, Y.; Xie, X. Molecular characteristics of natural organic matter in the groundwater system with geogenic iodine contamination in the Datong Basin, Northern China. *Chemosphere* **2023**, *333*, 138834. [[CrossRef](#)] [[PubMed](#)]
57. Xie, X.; Wang, Y.; Li, J.; Wu, Y.; Duan, M. Soil geochemistry and groundwater contamination in an arsenic-affected area of the Datong Basin, China. *Environ. Earth Sci.* **2014**, *71*, 3455–3464. [[CrossRef](#)]
58. Wu, Y.; Wang, Y.; Xie, X. Spatial occurrence and geochemistry of soil salinity in Datong basin, northern China. *J. Soils Sediments* **2014**, *14*, 1445–1455. [[CrossRef](#)]
59. Qian, K.; Sun, H.; Li, J.; Xie, X. Strontium isotopes as tracers for water-rocks interactions of groundwater to delineate iodine enrichment in aquifer of Datong Basin, northern China. *Appl. Geochem.* **2023**, *158*, 105783. [[CrossRef](#)]
60. Guo, H.; Wang, Y. Hydrogeochemical processes in shallow quaternary aquifers from the northern part of the Datong Basin, China. *Appl. Geochem.* **2004**, *19*, 19–27. [[CrossRef](#)]
61. Yi, Q.; Cheng, Y.p.; Zhang, J.k. Analysis on the salt content characteristics of southern saline-alkali soil in Datong Basin and its causes. *J. Groundw. Sci. Eng* **2014**, *2*, 63–72. [[CrossRef](#)]
62. Biedunkova, O.; Kuznietsov, P. Liquid Ion Chromatographic Determination of Soluble Ions in Water: Comparison of Greenness and Comprehensive Assessment of Irrigation Suitability. *Water Air Soil Pollut.* **2025**, *236*, 315. [[CrossRef](#)]
63. Gheyas, I.A.; Smith, L.S. Feature subset selection in large dimensionality domains. *Pattern Recognit.* **2010**, *43*, 5–13. [[CrossRef](#)]
64. Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction*; Springer New York: New York, USA, 2009.
65. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [[CrossRef](#)]
66. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
67. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
68. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
69. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
70. Hsu, C.W.; Chang, C.C.; Lin, C.J. *A Practical Guide to Support Vector Classification*; Technical report; Department of Computer Science, National Taiwan University: Taipei, Taiwan, 2003.
71. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [[CrossRef](#)]
72. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proceedings of the European Conference on Information Retrieval*, Santiago de Compostela, Spain, 21–23 March 2025; Springer: Berlin/Heidelberg, Germany, 2005; pp. 345–359.
73. Calle, M.L.; Urrea, V. Stability of Random Forest importance measures. *Briefings Bioinf.* **2011**, *12*, 86–89. [[CrossRef](#)]

74. Khalil, A.; Almasri, M.N.; McKee, M.; Kaluarachchi, J.J. Applicability of statistical learning algorithms in groundwater quality modeling. *Water Resour. Res.* **2005**, *41*, W05010. [[CrossRef](#)]
75. Ouedraogo, I.; Defourny, P.; Vanclooster, M. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrol. J.* **2019**, *27*, 1081–1098. [[CrossRef](#)]
76. Tesoriero, A.J.; Gronberg, J.A.; Juckem, P.F.; Miller, M.P.; Austin, B.P. Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. *Water Resour. Res.* **2017**, *53*, 7316–7331. [[CrossRef](#)]
77. Al-Mukhtar, M. Random forest, support vector machine, and neural networks to modelling suspended sediment in Tigris River-Baghdad. *Environ. Monit. Assess.* **2019**, *191*, 673. [[CrossRef](#)] [[PubMed](#)]
78. Adamowski, J.; Chan, H.F. A wavelet neural network conjunction model for groundwater level forecasting. *J. Hydrol.* **2011**, *407*, 28–40. [[CrossRef](#)]
79. Pi, K.; Wang, Y.; Xie, X.; Su, C.; Ma, T.; Li, J.; Liu, Y. Hydrogeochemistry of co-occurring geogenic arsenic, fluoride and iodine in groundwater at Datong Basin, northern China. *J. Hazard. Mater.* **2015**, *300*, 652–661. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.