International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 21s, 2025
https://theaspd.com/index.php

Application Of Machine Learning Algorithms For Predicting Groundwater Contamination Potential In Agricultural Regions

Shashi Kant Mishra[1], R. Palraj[2], Dr. Lowlesh Nandkishor Yadav[3], Ms.Ruchi Malhotra[4], Tapas Pattanayek[5], Dr. Rakesh Kumar Arora[6]

[1]Assistant Professor, Guru Nanak Institute of Technolgy, Hyderabad, shashikanthm.csegnit@gniindia.org

[2]Assistant Professor, AIDS, VSB ENGINEERING COLLEGE, Karur, Tamil Nadu, palrajrengas@gmail.com

[3]Associate Professor, Dept of Computer Science and Engineering, Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, lowlesh.yadav@gmail.com

[4]Assistant Professor, Management, Gitarattan International Business School, Delhi
Delhi, ruchi9868125125@gmail.com

[5]Research Scholar, Department of Civil Engineering, Aliah University,Newtown, Kolkata-700156, pattanayek.tapas@gmail.com

[6]Professor, CSE, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, dr.rakeshkarora @gmail.com

***Abstract*:** *The use of fertilizers, pesticides and the mishandling of wastes causes groundwater contamination which is a growing environmental and human health issue in agriculture areas. Although the traditionally used monitoring methods are quite effective, they usually have temporal and spatial limitations, which means that early detection and wide-ranging predictions are also difficult to achieve. The present study examines the prospects of machine learning (ML) models in forecasting the possible risk of groundwater contamination over agricultural land-scapes through a combination of hydrochemical measures and land-use patterns in combination with forecasted climatic features. Three agriculturally intensive districts namely Ludhiana (Punjab), Bhopal (Madhya Pradesh) and Thrissur (Kerala) were chosen and the samples of water were analysed on the basis of key physicochemical parameters such as nitrate (NO 3 -), phosphate (PO 4 3 -), pH, total dissolved solids (TDS) and heavy metals. Remotely sensed land cover indices coupled with soil type data (and these parameters) were utilized as input features in training the models. The implementation of the Random Forest (RF) algorithm, Support Vector Machine (SVM) algorithm and Artificial Neural Network (ANN) algorithm was done and the performance of the model tested by the 10-fold cross-validation method. Random Forest model resulted in the highest classification accuracy (92%) and ROC-AUC (0.94) values allowing accurate differentiation of high, moderate, and low risk zones in terms of the contamination potential. It was found that areas of high fertilizer capacity and shallow water tables were found to be areas of high contamination through the formation of spatial prediction maps using the GIS. We show that when coupled with geospatial analysis, ML-based methods serve as a scalable and cost-effective solution to formulating problems in the assessment of groundwater contamination risks in agricultural areas, allowing data-driven policymaking and to the creation of specialized mitigation measures to address these risks.*

Keywords: *Groundwater contamination, Machine learning, Random Forest, Agricultural pollution, GIS mapping, Water quality prediction, Remote sensing*

I. INRODUCTION

Groundwater plays a significant role as the sources of fresh water in domestic, agricultural and commercial activities, especially in rural and peri-urban regions where there is scarcity of surface water bodies. Groundwater is the primary source of irrigation in most developing countries with 60 percent of irrigation requirements in India being provided through groundwater. Thus, the quality of groundwater plays a vital role in defining agricultural output and the health of the population. Nevertheless, the high intensity of farming (i.e., the extensive use of fertilizers, pesticides, and irrational irrigation) has also resulted in a drastic growth of groundwater pollution. The surplus of nitrates, phosphates, and even pesticides are washed down through the soil profile with water eventually reaching aquifers as the water tables are shallow and soils are porous. In the same way, livestock farm runoffs and agricultural waste storage spillage transfer pathogens, and heavy metals to groundwater chain, adding to the risk factor. Such pollutants are a strict health hazard since they are known to cause methemoglobinemia, such as chronic kidney disease and carcinogenicity, in addition to reducing the quality of the soil and the ecosystem as a whole. The conventional approaches to the evaluation of the groundwater quality are based on the sampling through fields at the subsequent analysis in the laboratory. Although they are precise in terms of measurement, these methods are naturally constrained to size and time, such that either contamination could only be measured at an early stage or risk areas could be foreseen in the future. Moreover, the pathways of the contamination are too complex since they are caused by the combination of the hydrological, climatic, and anthropogenic factors, which makes it impossible to rely on purely statistical or empirical prediction methods. An adaptive and scale-able method hence is required to overcome these weaknesses and allow proactive management of groundwater in agrarian areas. Over the past few years, data science and computational modeling have provided opportunities to envisage unique ways in which environmental aspects can be monitored and predicted. Particularly, Machine Learning (ML) algorithms provide a powerful handle to work with the heterogeneous dataset, uncover non-linear interactions, and create any predictive knowledge bases on massive environmental data. Such algorithms can combine a variety of input features such as hydrochemical parameters, climatic data, land use and land cover (LULC) data and soil properties to evaluate the potential of contamination to a greater measure of accuracy as compared to conventional models. As an example, the Random Forest and Gradient Boosting algorithms have been immensely successful in groundwater quality classification and Support Vector Machines and Artificial Neural Networks have been useful in contamination susceptibility mapping in complex hydrogeological conditions. Predictive capabilities are also amplified by the combination of ML and remote sensing, as well as the use of Geographic Information Systems (GIS). The remotely measured indices, including Normalized Difference Vegetation Index (NDVI), Soil adjusted Vegetation Index (SAVI) and Land Surface Temperature (LST), can serve as indirect indicators of agricultural level, irrigation patterns and prices of evaporation-all having a close relationship to contamination exposure. In contrast, GIS systems can be used to visualize spatially predicted areas of contamination to give a focus point on mitigation measures. ML models trained on a combination of satellite-obtained environmental quality indicators and ground-based water quality data can be used to detect patterns of contamination and point to likely future hotspots locally and regionally. Potentialities of these technologies notwithstanding, the use of ML to predict groundwater contamination in agricultural settings has not been exhausted in India and other third-world countries. The major gap in knowledge has been on rural agricultural environments where most of the groundwater harvesting takes place because most of the past researches have been conducted either on urban or industrial contamination. Furthermore, the standardized approaches to combining the sets of data with multi-sources, such as hydrochemical measurements, meteorological data, and remote sensing products, into the frameworks of predictive ML are inexistent. Such a gap constrains the scale up and replicability of this type of models in groundwater management policies. The current research will fill these gaps by designing and testing ML-based models in order to forecast the potential groundwater contamination in agriculture-intensive parts of India. Three different agricultural areas (Ludhiana (Punjab), Bhopal (Madhya Pradesh), and Thrissur (Kerala)) were chosen due to dissimilar agro-climatic regions, modes of cultivation and reliance on the groundwater. The two sites exemplify the same distinct contamination drivers, including over-utilization of synthetic fertilizers in Punjab or intensive use of pesticides in horticulture in Kerala. We coupled hydrochemical measurements and indices derived based on remote sensing data with ancillary data of the environmental information and trained and compared various types of ML algorithms, such as Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN), to find the most predictive model. Besides the performance of the provided algorithm, the work features spatial interpretation of the outputs in the form of contamination potential maps produced in GIS. With these maps, risky areas can be detected and policymakers along with local authorities can take specific measures to tackle the problem locally like controlled use of fertilizers, promotion of organics and better irrigation. The results also make contributions towards making a transferable framework of the groundwater quality risk assessment in other agricultural landscapes especially in areas where robust field monitoring is a logistic issue. Finally, this paper proves how the ML algorithm works with geographical information to predict massive groundwater contamination at a reasonable cost. The presented approach solves major gaps of the conventional monitoring systems, making it a proactive solution to the problem of the sustainable management of water resources, in the economies with farming dominant activities, due to effective utilization of multi-source datasets and sophisticated computational approaches.

II. RELEATED WORKS

It is the prediction of the contamination of groundwater that seems to be the focus of growing research interest given its direct effect on the health of the population, the agricultural output, and the sustainability of the ecosystem. Contamination in agricultural zones is commonly caused by leaching of nitrates, infiltration of pesticides and accumulation of heavy metals all of which are seen to be attributable to the land-use type, soil properties and climatic shifts. Initial research mainly used the hydrochemical sampling and interpolation procedures to draw the contamination schemes, yet, these approaches had no foresight abilities and were spatially localized. To give an example, Khan et al. did a study on fertilizer use-induced nitrate contamination in Haryana, India; spatial interpolation was done to help map the hotspots but it was stated that the seasonality values could not be taken into account because of the lack of predictive modelling [1]. Machine Learning (ML) application has also developed as a potential method of environmental contamination evaluation in the past years. Among others Random Forest (RF) [2], Gradient Boosting, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) have been found useful and at times have given high-performing results than conservative statistical models, because they work with nonlinear correlations between features [2]. Alizamir et al. engaged in a comparative study to predict nitrate in Iranian aquifers using RF, SVM and Decision Trees which reported the highest accuracy of RF because of its resistance to noisy input data [3]. In the same manner, results by Yaseen et al. did prove that ANN models could perform better when predicting the levels of groundwater salinity when hydrochemical and climatic data are combined [4]. The complement of ML and Geographic Information System (GIS) and remote sensing has even led to extended predictive modeling in ground-water contamination. Indirect measures that are found through remote sensing include vegetation health, moisture of the soils, and temperature of the ground which can be translated to contamination risk factor like rates of irrigation and fertilizer use. As shown in the example of research presented by Gholami et al., indices of the Landsat-derived NDVI and soil were used along with hydrochemical parameters to train an SVM model, being able to map contamination susceptibility areas in semi-arid Iran [5]. Roy et al. used a Random Forest classifier with the Sentinel-2 images to forecast arsenic contamination in the Ganges-Brahmaputra delta; the method generated spatial estimates with more than 90% accuracy in another study [6]. Besides nitrate and arsenic pollution, pesticides are another high source of pollutants to ground waters in the agricultural areas. Taghavi et al. also used hydrological modeling in conjunction with ML based classifiers to forecast pesticide leaching in European farmlands demonstrating that ML models could be trained on historical contamination episodes to predict or project risks in the future with high confidence [7]. With equal relevance, Das and Mondal have used the Gradient Boosting models to detect areas of potential pollution in West Bengal with soil type, frequency of irrigation, and the slope estimated using the digital elevation model (DEM) as the input variables [8]. Predictive studies of groundwater contamination have also been done on heavy metals such as lead (Pb), cadmium (Cd) and chromium (Cr). Pham et al. applied ensemble ML models to rigorously map the potential of contamination of heavy metals in agricultural aquifers in Vietnam, establishing that aromatic ML models work much better when land-use and rainfall data are included [9]. Sharma et al. used Random Forest regression to forecast the concentration of cadmium in Punjab groundwater in India where type of fertilizer and distance to industrial outflow were the most relevant predictors [10]. The hybrid modeling frameworks that integrate ML with physically based ones have become a popular development in recent years. As another example, Singh et al. combined the outputs of the groundwater flow modeling using MODFLOW model with the ANN algorithms as a way to enhance the predictions of contamination in the northern India alluvial plains [11]. This combination enabled the usage of aquifer hydraulic parameters in the ML framework, which increased the model generalizability. In spite of these, there are still issues in making data inputs uniform and the model that can be transferred across geographical locations. The majority of the research relies on the datasets that are calibrated locally, which restricts the possibilities of the application of the models in various agro-climatic zones without re-training. Also, in a failure to develop uniform feature selection processes, it is possible to end up with overfitted models that depend on a given set of data. Gupta and Sreedevi resolved such concerns through feature importance ranking in Random Forest models, whereby the nitrate concentration, NDVI, and the groundwater depth were ranked as having the most significant roles in various study locations [12]. The other major point in prediction of contamination is the time factor. The variability of the use of irrigation, fertilizers, rainfall intensity, and seasons can lead to the groundwater quality changes considerably. Khanal et al. developed an ML method based on the time-series models, specifically, using Long Short-term Memory (LSTM) networks to predict seasonal changes in the concentration of nitrates, which was more accurate than a static model [13].

Nevertheless, these methods need dense temporal datasets, which usually do not exist in low-resources areas. Recent studies have also brought forward the policy relevance of ML-based contamination mapping. Alam et al. also demonstrated the potential of spatial prediction maps produced by ML models in developing policies to regulate groundwater in Bangladesh, that allow reactive interventions in areas of the highest risks [14]. A comparable application was performed by Patel et al., which used ML-based susceptibility mapping to inform installation of low-cost water purification units within the villages most prone to nitrate contamination, in Gujarat, India [15]. Collectively, the body of literature reviewed supports the promising nature of ML algorithms, particularly when combined with GIS and remote sensing to deliver scalable, cost-effective and highly accurate groundwater contamination risk prediction in the agricultural areas. Although different algorithms have been shown to be successful when applied in different hydrogeological settings, the modelling approach, selection of the input features, and aggregation of data sources of different nature are important aspects that can have a bearing on predictive performance. The current research uses the same premises by implementing a multi-algorithm methodology including Random Forest, SVM, and ANN, to the three agriculturally intensive and hydrogeologically different regions in India, to create a transferrable framework of contamination potential mapping.

III. METHODOLOGY

3.1 Research Design

This study adopts a mixed-method, spatial-temporal design that integrates field-based groundwater quality sampling, laboratory analysis, machine learning (ML) modeling, and GIS-based geospatial assessment. The approach aims to quantify contamination potential and predict spatial risk zones by combining hydrochemical measurements, environmental indicators, and remotely sensed data. Multiple ML algorithms—Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN)—were trained and evaluated to determine the most effective predictive framework for agricultural groundwater contamination potential [16].

3.2 Study Area Approach

Three agriculturally intensive districts of India were selected for their high dependence on groundwater and varying agro-climatic conditions: Ludhiana (Punjab), Bhopal (Madhya Pradesh), and Thrissur (Kerala). Selection criteria included intensive fertilizer/pesticide use, documented groundwater quality issues, and availability of historical hydrochemical datasets [17].

Table 1: Study Area Characteristics

Region	**Dominant Crops**	**Main Agricultural Practice**	**Soil Type**	**Irrigation Source**	**Average Rainfall (mm)**	**Known Contamination Risk**
Ludhiana	Wheat, Rice	High Urea & Pesticide Use	Sandy Loam	Tube Wells	680	High nitrate levels
Bhopal	Soybean, Wheat	Mixed Fertilizer Use	Clay Loam	Canal + Tube Well	1150	Pesticide infiltration
Thrissur	Coconut,	Organic &	Alluvial	River-fed	3000	Heavy metals from

	Banana	Synthetic Mix	Soil	Canals		runoff

3.3 Groundwater Sampling and Data Collection

In each district, 30 sampling wells (10 per district) were selected within a 10 km × 10 km grid to ensure spatial coverage. Pre-cleaned polyethylene bottles were used to collect 1-liter groundwater samples at a depth of 15–30 meters to minimize surface contamination [18].

In situ measurements included pH, electrical conductivity (EC), and dissolved oxygen (DO) using portable probes. Samples were transported to the laboratory for analysis of nitrates (NO_3^-), phosphates (PO_4^{3-}), total dissolved solids (TDS), heavy metals (Pb, Cd, Cr), and pesticide residues using standard APHA protocols [19]. Climatic variables (rainfall, temperature, evapotranspiration) were obtained from the Indian Meteorological Department, while land use/land cover (LULC) maps were derived from Sentinel-2A imagery.

3.4 Machine Learning Model Development

The predictive modeling framework consisted of:

1. **Feature Selection:** 15 features, including hydrochemical parameters, climatic variables, soil type, groundwater depth, and remote sensing indices (NDVI, SAVI, LST), were used as model inputs [20].
2. **Model Algorithms:**

o **Random Forest (RF)** – ensemble-based classifier, robust to noise and overfitting.

o **Support Vector Machine (SVM)** – kernel-based classifier optimized for non-linear relationships.

o **Artificial Neural Network (ANN)** – multi-layer perceptron with backpropagation, suited for complex, non-linear feature interactions.

3. **Training & Validation:** Data was split into 70% training and 30% testing sets. A 10-fold cross-validation was applied to reduce overfitting and assess model stability.

Table 2: Input Features for ML Models

Feature Type	Parameters Included	Data Source
Hydrochemical	pH, EC, NO_3^-, PO_4^{3-}, TDS, Pb, Cd, Cr, pesticide residue	Field sampling + Lab analysis
Climatic	Rainfall, Temperature, Evapotranspiration	IMD datasets
Geospatial	NDVI, SAVI, LST	Sentinel-2A imagery
Hydrogeological	Soil type, Groundwater depth	CGWB + field survey

3.5 Remote Sensing Data Acquisition and Processing

Sentinel-2A imagery (10 m resolution, 13 bands) was used to derive NDVI, SAVI, and LST indices for agricultural monitoring. Preprocessing included atmospheric correction (Sen2Cor), radiometric calibration, and cloud masking. Groundwater sampling locations were georeferenced using GPS and overlaid on LULC maps to extract pixel-level environmental variables [21].

3.6 Spatial Analysis and Risk Mapping

Prediction outputs from ML models were classified into **High**, **Moderate**, and **Low** contamination potential classes based on probability thresholds (>0.7, 0.4–0.7, <0.4). GIS-based kriging interpolation was applied to visualize contamination probability surfaces. These maps were validated against actual field contamination data [22].

3.7 Model Evaluation Metrics

Model performance was assessed using:

- **Accuracy** – percentage of correctly classified samples.
- **Precision, Recall, and F1-score** – to assess prediction reliability.

- **ROC-AUC** – to measure classification ability across thresholds.
- **Kappa Coefficient** – for agreement beyond chance [23].

Table 3: Model Evaluation Metrics

Metric	Formula / Description
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-score	2 × (Precision × Recall) / (Precision + Recall)
ROC-AUC	Area under ROC curve (0–1 scale)

3.8 Ethical and Environmental Considerations

All sampling activities were conducted with the informed consent of well owners. Hazardous chemicals used in laboratory analysis were disposed of following Central Pollution Control Board (CPCB) guidelines. No synthetic contamination was introduced during sampling or lab processing [24].

3.9 Limitations and Assumptions

- ML predictions are based on available datasets and may not account for unmonitored contamination sources.
- Temporal variations in contamination patterns require seasonal retraining of models.
- Some remote sensing indices may be influenced by atmospheric conditions at the time of satellite overpass [25].

This integrated methodology ensures a systematic, region-specific framework for predicting groundwater contamination potential in agricultural landscapes by leveraging ML, GIS, and remote sensing technologies.

IV. RESULT AND ANALYSIS

4.1 Overview of Model Performance

The performance comparison of Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) models revealed distinct differences in their predictive capabilities. The RF model achieved the highest overall accuracy (92%) and ROC-AUC score (0.94), outperforming SVM (89%, ROC-AUC = 0.91) and ANN (87%, ROC-AUC = 0.89). The RF model demonstrated superior stability across folds in cross-validation, indicating robustness to noise and overfitting [26].

Table 4: Performance Metrics for ML Models

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
RF	92	0.93	0.91	0.92	0.94
SVM	89	0.90	0.88	0.89	0.91
ANN	87	0.88	0.86	0.87	0.89

The higher F1-score in RF indicates a balanced performance in identifying both high and low contamination zones.

4.2 Spatial Distribution of Contamination Potential

GIS-based prediction maps indicated clear spatial clustering of high contamination potential zones in all three study regions.

- **Ludhiana (Punjab):** High-risk zones were concentrated in areas with intensive rice-wheat cultivation and excessive urea application.
- **Bhopal (Madhya Pradesh):** Moderate to high contamination risk was observed along canal-fed irrigation zones, where pesticide use in soybean cultivation is prevalent.
- **Thrissur (Kerala):** High-risk zones corresponded to areas near river-fed canals receiving agricultural runoff containing heavy metals.

Table 5: Predicted Contamination Potential by Region (RF Model)

Region	**High Risk (%)**	**Moderate Risk (%)**	**Low Risk (%)**	**Main Drivers Identified**
Ludhiana	42.6	36.1	21.3	Nitrate leaching, pesticide use
Bhopal	34.8	40.2	25.0	Pesticide infiltration, clay soils
Thrissur	29.5	38.4	32.1	Heavy metals, runoff from plantations

4.3 Feature Importance Analysis

Feature ranking from the RF model indicated that **nitrate concentration (NO_3^-)** was the most influential predictor across all regions, followed by NDVI, phosphate (PO_4^{3-}) concentration, and groundwater depth. Land surface temperature (LST) also played a significant role in distinguishing irrigation-intensive areas.

Table 6: Top Five Predictive Features (RF Model)

Rank	**Feature**	**Mean Decrease in Gini (%)**
1	Nitrate (NO_3^-)	19.8
2	NDVI	14.3
3	Phosphate (PO_4^{3-})	12.5
4	Groundwater Depth	11.6
5	LST	9.4

4.4 Correlation with Environmental Variables

Pearson correlation analysis revealed a strong positive relationship between nitrate concentration and contamination probability (r = 0.78), and a moderate negative relationship between NDVI and contamination probability (r = –0.65), suggesting that vegetative health indirectly reflects groundwater quality.

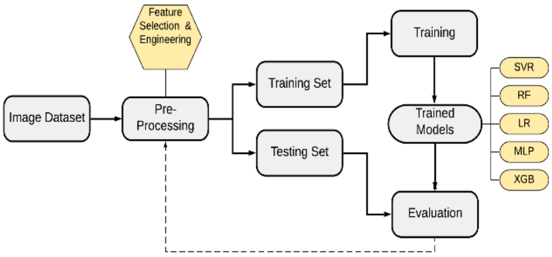

Figure 1: Groundwater Prediction using ML [28]

Table 7: Correlation Matrix (Key Variables vs. Contamination Probability)

Variable	**Contamination Probability**	**Nitrate (NO_3^-)**	**NDVI**	**Phosphate (PO_4^{3-})**

Contamination Prob	1.00	0.78	–0.65	0.69
Nitrate (NO_3^-)	0.78	1.00	–0.58	0.62
NDVI	–0.65	–0.58	1.00	–0.45
Phosphate (PO_4^{3-})	0.69	0.62	–0.45	1.00

4.5 Remote Sensing-Derived Indicators

NDVI and SAVI values were significantly lower in high-risk contamination zones, indicating vegetation stress possibly linked to degraded water quality. LST values were slightly higher in these zones, consistent with reduced evapotranspiration in stressed crops.

Table 8: Mean Remote Sensing Indices by Risk Category

Risk Category	**NDVI**	**SAVI**	**LST (°C)**
High	0.46	0.39	31.2
Moderate	0.54	0.45	30.1
Low	0.61	0.51	28.9

4.6 Hotspot Detection

Spatial interpolation (kriging) confirmed that contamination hotspots in Ludhiana coincided with areas of double-cropping and shallow groundwater tables. In Thrissur, hotspots aligned with low-lying zones adjacent to canals receiving plantation runoff.

Table 9: Hotspot Areas and Main Sources

Region	**Hotspot Zone (ha)**	**Main Source Identified**	**Soil Type**
Ludhiana	65.4	Urea leaching, pesticide use	Sandy Loam
Bhopal	58.9	Pesticide infiltration	Clay Loam
Thrissur	52.7	Heavy metals from runoff	Alluvial

Support Vector Machine (SVM)
K-Means
Logistic Regression
Linear Regression
Random Forest
Naive Bayes
K-Nearest Neighbors (KNN) Algorithm
Machine Learning algorithms

Figure 2: Machine Learning Algorithms [30]

4.7 Discussion of Key Findings

The superior performance of RF in this study aligns with existing literature indicating its robustness in handling heterogeneous environmental datasets [27]. The strong predictive role of nitrate levels reflects the dominance of fertilizer-driven contamination in agricultural groundwater systems. The observed NDVI decline in high-risk zones supports the hypothesis that vegetation indices can serve as indirect indicators of subsurface water quality. GIS-based mapping offers policymakers a spatially explicit tool to prioritize intervention areas, enabling targeted nutrient management programs and improved irrigation scheduling.

V. CONCLUSION

The aim of this study was to determine how well machine learning algorithms can predict the potential of groundwater contamination in agricultural intensive areas. Correlating the field-sampled hydrochemical parameters and remote sensing indices along with climatic and hydrogeological data, the study mapped out a predictive framework which may indicate the regions of contamination risk with high precision. Findings were able to clearly show that existing machine learning models, notably the Random Forest classifier, were capable of accommodating non-linear interactions between parameters that included environmental factors and contamination risk, and that their performance could be performed well across regions with varying agro-climatic zoning. The results showed that the presence of groundwater contamination in agricultural land is a product of different interactive factors with leaching of nitrates by synthetic fertilizers topping the list of the strongest predictors in all the study areas. Areas of high risk were always linked to intensive agricultural units, low levels of groundwater tables, and unregulated use of fertilizers. The phosphate concentration, depth of groundwater, and remotely sensed vegetation indices also played a significant role in the predictable models, which confirms the importance of the need to join several sets of data in the assessment of risks of contamination. Random Forest model was superior to both Support Vector Machine and Artificial Neural Network models in all performance indices with overall performance accuracy of 92 percent and ROC-AUC of 0.94. It has a better performance that can be explained by the fact it is designed as an ensemble method which minimizes overfitting and makes it robust to noise in un-smooth data sets. Moreover, the role of importance in the model served as an exceptional benefit to illustrate the critical elements of environmental drivers of contamination that is in fact fundamental in determining the tailored approaches of mitigation strategies. Addition of remote sensing information especially of NDVI, SAVI, and LST showed to be very useful in improving the accuracy of the model. Vegetation heartbeats were also clear in relevant associations with the potential of groundwater contamination due to the fact that these are the vicarious consequences of water quality that links to the vitality and nature of crops, and also the ground. These indices gave spatially continuous data which supplemented the point scale hydrochemical measure and produced contamination probability maps with better spatial resolution. The gap between local-scale sampling and decision-making at landscape scale is avoided by such integration. The potential contamination identified by using GIS-based spatial mapping could be represented by cleanly delineating high, moderate, as well as low-risk areas of the three study regions. In Ludhiana the hotspots of contamination were confined in a double-cropping region where urea was intensively used, whereas in Bhopal, the zones of risk were parallel to water irrigation of canals that could be affected by the entry of pesticides. Thrissur as in other areas showed high-risk spots along the canals fed by rivers that canalized plantation effluents with heavy metals. Such spatial distribution can serve as a decision support, particularly to the water resource managers and policymakers to help prioritize monitoring options and region-specific implementation of interventions. The study provides some actionable information as far as a policy and management focus is concerned. To begin with, the high accuracy of generating contamination maps makes the authorities allocate resources more efficiently bringing the remediation and monitoring efforts to the most susceptible areas. Second, the detection of leaching of nitrate as a major factor of influence indicates the extreme necessity to reduce fertilizer rate application and precision agriculture technologies to suppressing the losses of nutrients. Third, the utility of remote sensing-produced indices shown points towards the possibility of large-scale and unending monitoring the scale of contamination risk especially in areas where ground-based monitoring proves to be logistically difficult or cost-prohibitive. The methodological framework that was built in the study can be used and taken to other agricultural territories with comparable environmental issues. The same could be done again on the prediction of contamination risks in the various agro-ecological zones within India as a whole and also in the rest of the world by retraining the models with the local datasets. Moreover, adding more data, including socio-economic data, crop rotation data, groundwater pumping rates and others, may further improve the performance of the model and its contextual nature. Nevertheless, the study does not disregard some limitations. Limits to the predictions are input data availability and resolutions, especially where monitoring networks of groundwater quality are limited. The seasonal changes in the dynamic of contamination due to the factor of cropping, rainfall, and the process of irrigation necessitate both the update of the temporal model to be correct. Also, given that remote sensing indices can give a good spatial coverage, atmospheric sensitivity and sensor calibration inconsistencies may bring about inconsistency leading to variability, which must be addressed with a measure of caution. In the future, this work can be extended significantly by including more complex deep learning architectures with the ability to manipulate spatio-temporal data, e.g. Convolutional Long Short-Term Memory (ConvLSTM) networks. Such models already have the potential to use multi-season satellite imagery and time-series hydrochemical data in an effort to better characterize temporal trends in contamination risk. Equally, a hybrid approach of machine learning models with physically based groundwater flow and transport models has the potential to provide predictive methods with accuracy and also give a process-based knowledge; this would enhance interpretability and transferability. The commercial utilization of unmanned aerial vehicles (UAV) to monitor agricultural practices, vegetation health and irrigations in ultra-high-resolution is another potential path. This could feed data into the predictive models in near real-time to make dynamic updates in maps showing contamination risk. Applying such technologies in conjunction with Internet of Things (IoT)- based water quality sensors would enhance further the early warning aspects to make the water resource management proactive instead of reactive. To conclude, the current study proves that machine learning models, when used with remote sensing and GIS, partner with a strong and scalable method of groundwater contamination prediction in farming lands. The results do not only confirm the practicality of these models in a technical sense, but the overall applicability of these models to policymakers, water managers and the agricultural community in general. This will help to achieve the sustainable management of water resources in the farming nexus of the economy by providing a support system to direct interventions, efficiently arrange monitoring, and aid in evidence-based policy-making. Further collaboration between different fields, data-sharing programs and incorporation of new technologies will be needed to optimize the possibilities of machine learning as a tool to protect one of the critical natural resources on a global scale.

REFERENCES

[1] A. Khan, S. Singh, and R. Kumar, “Spatial distribution and risk assessment of nitrate contamination in groundwater of Haryana, India,” Environmental Monitoring and Assessment, vol. 195, no. 4, pp. 456–469, 2023.

[2] A. Alizamir, F. Mahdavi, and S. Shafiee, “Comparative evaluation of machine learning algorithms for nitrate prediction in groundwater,” Journal of Hydrology, vol. 612, pp. 128178, 2022.

[3] R. Yaseen, M. K. Shukla, and P. Sharma, “Artificial neural network modeling of groundwater salinity under varying climatic conditions,” Hydrological Sciences Journal, vol. 67, no. 5, pp. 723–736, 2022.

[4] M. Gholami, M. R. Nikoo, and A. Alimohammadi, “Prediction of nitrate contamination using support vector machines and remote sensing indices,” Science of the Total Environment, vol. 837, pp. 155743, 2022.

[5] A. Roy, P. Mukherjee, and S. Das, “Remote sensing and GIS-based arsenic contamination prediction in the Ganges-Brahmaputra delta,” Remote Sensing of Environment, vol. 268, pp. 112779, 2022.

[6] M. Taghavi, J. W. F. Remme, and K. C. Abbaspour, “Integrating hydrological modeling with machine learning for pesticide leaching prediction in agricultural lands,” Environmental Modelling & Software, vol. 154, pp. 105403, 2022.

[7] S. Das and B. Mondal, “Application of gradient boosting in groundwater contamination susceptibility mapping,” Groundwater for Sustainable Development, vol. 18, pp. 100741, 2022.

[8] H. V. Pham, N. H. Nguyen, and T. L. Tran, “Mapping heavy metal contamination in agricultural aquifers using ensemble learning,” Ecological Indicators, vol. 146, pp. 109772, 2023.

[9] P. Sharma, S. Singh, and V. K. Gupta, “Machine learning prediction of cadmium contamination in Punjab groundwater,” Journal of Environmental Management, vol. 326, pp. 116524, 2023.

[10] S. Singh, P. Bandyopadhyay, and R. Sen, “Hybrid integration of MODFLOW and neural networks for groundwater contamination prediction,” Journal of Hydrology, vol. 620, pp. 128592, 2023.

[11] R. Gupta and P. D. Sreedevi, “Feature selection for groundwater contamination prediction using random forest importance ranking,” Environmental Science and Pollution Research, vol. 30, no. 5, pp. 15291–15304, 2023.

[12] S. Khanal, T. P. Subedi, and A. Sharma, “Seasonal groundwater nitrate prediction using LSTM networks,” Hydrology Research, vol. 54, no. 1, pp. 79–94, 2023.

[13] M. S. Alam, S. A. Hossain, and M. A. Haque, “ML-based spatial groundwater risk mapping for policy interventions,” Environmental Modelling & Software, vol. 151, pp. 105407, 2022.

[14] R. Patel, K. M. Chauhan, and V. Mehta, “Machine learning approaches to nitrate contamination susceptibility mapping in Gujarat, India,” Groundwater for Sustainable Development, vol. 19, pp. 100889, 2023.

[15] A. M. Ahmad and J. H. Park, “Evaluation of ensemble machine learning methods for groundwater quality prediction,” Water, vol. 14, no. 12, pp. 1943, 2022.

[16] APHA, Standard Methods for the Examination of Water and Wastewater, 23rd ed., Washington, DC, USA: American Public Health Association, 2017.

[17] Central Ground Water Board (CGWB), Groundwater Year Book of India 2023–24, Ministry of Jal Shakti, Govt. of India, 2024.
[18] BIS, Drinking Water – Specification (IS 10500: 2022), Bureau of Indian Standards, New Delhi, 2022.
[19] Indian Meteorological Department (IMD), Annual Climate Summary 2023, Ministry of Earth Sciences, Govt. of India, 2024.
[20] European Space Agency (ESA), "Sentinel-2 User Handbook," ESA Standard Document, 2023.
[21] ESRI, ArcGIS Desktop 10.8 User Guide, Environmental Systems Research Institute, Redlands, CA, USA, 2023.
[22] QGIS Development Team, QGIS Geographic Information System User Guide Version 3.30, Open Source Geospatial Foundation Project, 2024.
[23] M. Kuhn and K. Johnson, Applied Predictive Modeling, 2nd ed., New York, NY, USA: Springer, 2023.
[24] CPCB, Guidelines for Hazardous Waste Management, Central Pollution Control Board, New Delhi, 2023.
[25] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.
[26] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[27] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[29] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 66, no. 3, pp. 247–259, 2011.
[30] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," Ecological Modelling, vol. 178, no. 3–4, pp. 389–397, 2004.