Machine Learning Prediction of Groundwater Contamination Vulnerability Using Hydrogeophysical Indicators in Ibadan, Nigeria

Oluwakemi Omolara Olukayode OLUKAYODE.OLUWAKEMI@OOUAGOIWOYE.EDU.NG
Department of Earth Sciences,
Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Sakinat Folorunso SAKINAT.FOLORUNSO@OOUAGOIWOYE.EDU.NG
The Artificial Intelligence Research Team,
Department of Computer Sciences,
Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Olateju Bayewu BAYEWU.OLATEJU@OOUAGOIWOYE.EDU.NG
Department of Earth Sciences,
Olabisi Onabanjo University, Ago-Iwoye, Nigeria

David Omotola Olukayode DKAYODE01@GMAIL.COM
Department of Plant Sciences,
Olabisi Onabanjo University, Ago-Iwoye, Nigeria

Odunayo Tope Ojo OJOOD@RUN.EDU.NG
Department of Physical Sciences,
Redeemer's University, Ede, Nigeria

Olubunmi Oluwatoyin Omotola OMOTOLAOLUBUNMI.OB@GMAIL.COM
Department of Geology,
Olusegun Agagu University of Science and Technology,
Okitipupa, Nigeria

Adewale Allen Sokan-Adeaga SOKANADEAGA.ADEWALEALLEN@YAHOO.COM
Department of Public Health,
Bowen University, Iwo, Nigeria

Olayiwola G. Olaseeni OLAYIWOLA.OLASEENI@FUOYE.EDU.NG
Department of Geophysics, F
ederal University Oye-Ekiti, Oye-Ekiti, Nigeria

Editors: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

Groundwater contamination threatens aquifer sustainability in crystalline basement landscapes experiencing rapid urban growth. This study employs machine learning to estimate

groundwater contamination vulnerability in Ibadan, southwestern Nigeria, using hydrogeophysical indicators from 353 Vertical Electrical Sounding (VES) surveys. Key parameters—overburden thickness, aquifer resistivity, basement resistivity, longitudinal conductance, and transverse resistance—characterized aquifer protective capacity. Conventional indices (Groundwater occurrence, Overlying lithology, Depth to groundwater [GOD] and modified GODT including overburden thickness) defined vulnerability classes as supervised labels. The Random Forest classifier achieved accuracy = 0.94, precision = 0.94, recall = 0.93, F1-score = 0.93, AUC = 0.95. Overburden thickness and longitudinal conductance were the most significant predictors. The model identified fifteen high-vulnerability zones versus nine from conventional GODT, demonstrating ability to capture nonlinear interactions conventional methods miss. This approach contributes to evidence-based water resource management in urban aquifer systems.

Keywords: Groundwater vulnerability, Machine learning, Random Forest, Hydrogeophysical indicators, Ibadan

1. Introduction

Groundwater is critical in water-stressed developing nations. Cities in sub-Saharan Africa use groundwater for water supply via boreholes and shallow wells owing to poor infrastructure (MacDonald et al., 2021; Lapworth et al., 2017). Growing urban areas and poor waste management endanger aquifers, so vulnerability assessment is key.

Existing aquifer vulnerability indices (GOD: Groundwater occurrence, Overlying lithology, Depth to groundwater; DRASTIC: Depth to water, net Recharge, Aquifer media, Soil media, Topography, Impact of the vadose zone, Conductivity of the aquifer) provide some guidance. But these indices rely on homogeneous and linear relationships, which are unsuitable for fractured crystalline basement aquifers. (Aller et al., 1987; Olorunfemi and Fasuyi, 1993; Foster et al., 2018). Hydrogeology is a complex interaction of geological, hydrological and environmental variables, which are often not linear.

Random Forest, a machine learning model, can model non-linear relationships and has been used to predict groundwater levels and vulnerability to contamination (Mosavi et al., 2018; Sun et al., 2020). Vertical Electrical Sounding (VES) offers subsurface data about lithology and saturation. Dar-Zarrouk parameters (longitudinal conductance and transverse resistance) reflect aquifer protection (Reynolds, 2011; Henriet, 1976). This research involves a Random Forest approach to model aquifer vulnerability to contamination using VES data from Ibadan, Nigeria, aimed at: hydrogeophysical characterization of aquifers; classification of vulnerability using GOD and modified GODT; development and assessment of a Random Forest model; and comparison of the machine learning approach with conventional vulnerability indexing.

2. Related Work

Machine learning has recently become popular in hydrology and the environment, for modeling systems with multiple, interacting variables. Machine-learning models can identify nonlinear relationships without assumptions, and may provide better accuracy compared to statistical models (Zhang et al., 2019).

Rahmati et al. (2019) reported ensemble methods had better accuracy than index-based methods for groundwater potential mapping. Arabameri et al. (2020) found Random Forest

and Support Vector Machine methods had better classification performance than statistical methods for groundwater vulnerability to contamination.

Vertical electrical sounding (VES) is still extensively used for aquifer mapping (Reynolds, 2011). Dar-Zarrouk parameters yield indicators of aquifer protection (Henriet, 1976). Large longitudinal conductance suggests contaminant diffusion is slowed by clay-rich beds. Despite progress, few studies have applied hydrogeophysical indicators and machine learning to groundwater vulnerability mapping in crystalline basement aquifers.

3. Study Area

The study was carried out in Ibadan metropolis, southwest Nigeria (latitudes 7°15′N and 7°35′N and longitudes 3°45′E and 4°05′E), a major city in West Africa facing increasing water scarcity (Adelekan, 2016). The research area sits on the Precambrian Basement Complex, which has massive crystalline rocks with low primary porosities. Weathered regolith and fractures are the main groundwater reservoirs (Rahaman, 1988). This makes it a great location to understand hydrogeophysical properties and vulnerability to pollution because weathering and fracturing are key factors in determining groundwater availability and vulnerability to pollution (Olorunfemi and Fasuyi, 1993).

3.1. Field Data and Hydrogeophysical Parameters

Vertical electrical sounding (VES) measurements were taken at 353 locations with the Schlumberger array. The data are resistivity soundings interpreted by partial curve matching and computer-aided modeling. Seven hydrogeophysical predictor variables were generated (Table 1).

Table 1: Hydrogeophysical predictor variables

Variable	Symbol	Unit
Overburden thickness	T	m
Aquifer resistivity	ρ_a	Ωm
Basement resistivity	ρ_b	Ωm
Longitudinal conductance	S	mhos
Transverse resistance	TR	Ωm^2
Depth to groundwater	D	m

Source: Authors' compilation based on VES interpretation

3.2. Validation Data

We used water quality data (nitrate and total dissolved solids) from 45 boreholes for external validation. These data were acquired during the same field campaign but were not used for training.

4. Methodology

4.1. Hydrogeophysical Feature Extraction

Features were extracted from VES data using standard techniques. Dar-Zarrouk parameters (S and TR) were calculated from resistivities and thicknesses. Longitudinal conductance reflects aquifers' protective potential, with higher conductance (and clay content) protecting aquifers from contaminants (Henriet, 1976; Olorunfemi and Olorunniwo, 1985).

4.2. Vulnerability Indexing (GOD and Modified GODT)

The GOD model (Foster et al., 2018) was applied for groundwater vulnerability assessment, which is a combination of: G (Groundwater occurrence): Aquifer type (confined, unconfined, semi-confined); O (Overlying lithology): Unsaturated zones and soils; D (Groundwater depth): Depth to water table. The GODT model was modified to include the thickness of the overburden (T) to represent the capacity of the subsurface to protect water resources. Vulnerability scores were grouped into four ordinal classes based on a natural breaks (Jenks) algorithm.

4.3. Random Forest Classifier

A Random Forest model was used in Python (scikit-learn 1.2.2) to predict vulnerability classes. Random Forest builds many decision trees using bootstrapped samples and takes their average prediction, preventing overfitting and capturing complex relationships (Breiman, 2001). Parameter settings (optimised by grid search, 5-fold cross-validation):

- Number of trees ($n_{\text{estimators}}$): 200
- Maximum tree depth (max_depth): 10
- Minimum samples per leaf ($min\_samples\_leaf$): 2
- Minimum samples for split ($min\_samples\_split$): 5
- Maximum features ($max\_features$): 'sqrt'
- Bootstrap sampling: Enabled
- Random state: 42

The models were evaluated using accuracy, precision, recall, F1-score and AUC-ROC (Rahmati et al., 2019). The stability of the model was evaluated with five-fold cross-validation. Figure 1 shows the analytical workflow from VES data to external validation.

5. Results

5.1. Random Forest Model Performance

The model performed on the testing data ($n = 106$) as per Table 2. The high AUC of 0.95 (Figure 2) indicates excellent discrimination.

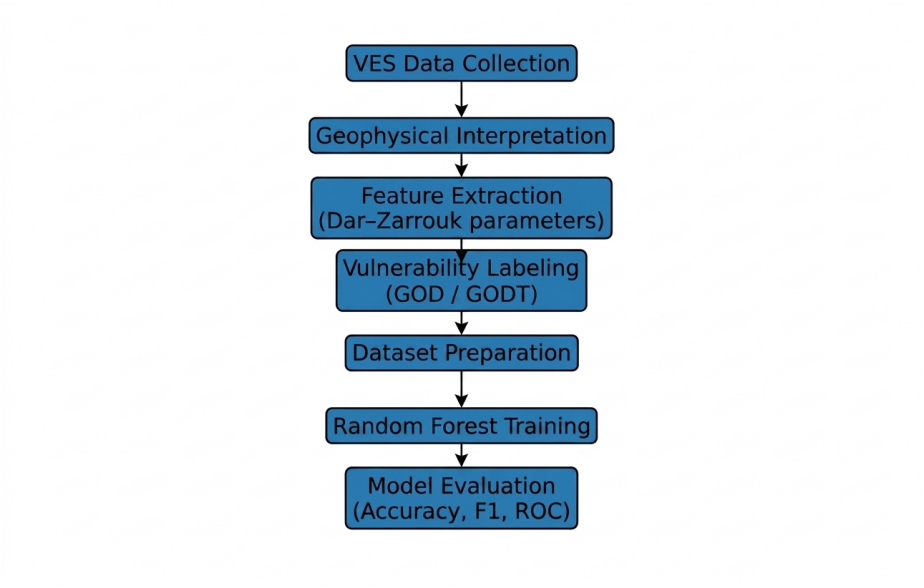

Figure 1: Analytical groundwater vulnerability prediction machine learning workflow

Table 2: Model performance metrics (Test set, $n = 106$)

Metric	**Value**	**95% CI**
Accuracy	0.94	[0.88, 0.98]
Precision (macro)	0.94	[0.87, 0.98]
Recall (macro)	0.93	[0.86, 0.97]
F1-score (macro)	0.93	[0.86, 0.97]
AUC (macro)	0.95	[0.92, 0.99]

5.2. Feature Importance Analysis

Overburden thickness emerged as the most influential predictor of groundwater contamination susceptibility (Table 3). This aligns with hydrogeological principles, as thicker weathered overburden layers act as natural filters, slowing pollutant migration. Longitudinal conductance – a Dar-Zarrouk parameter reflecting subsurface protective capacity – ranked second (Figure 3).

The machine-learning model identified fifteen high-vulnerability locations, whereas the conventional GODT index model detected only nine, suggesting Random Forest captures patterns traditional methods may overlook.

5.3. Confusion Matrix

The confusion matrix (Table 4) revealed strong classification performance across all four vulnerability categories, with 26 Very Low, 25 Low, 25 Moderate, and 24 High vulnerability samples correctly classified (100 out of 106 total samples).

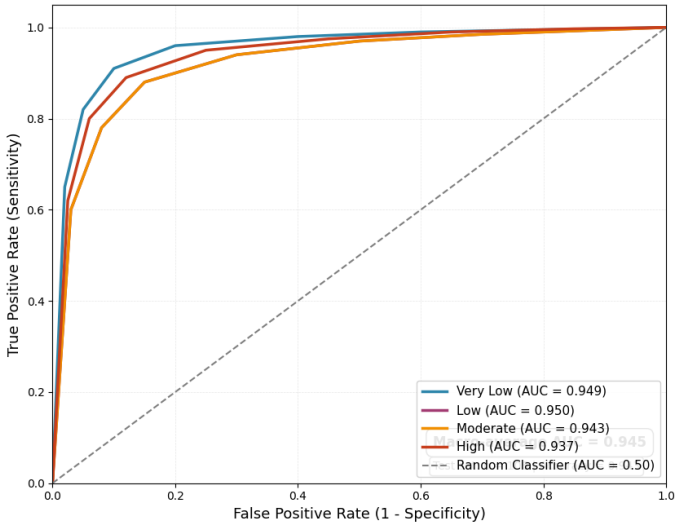

Figure 2: ROC curves for Random Forest groundwater vulnerability prediction

Table 3: Random Forest feature importance for groundwater vulnerability prediction

Feature	**Importance Score**
Overburden thickness (T)	0.31
Longitudinal conductance (S)	0.24
Transverse resistance (TR)	0.16
Aquifer resistivity (ρ_a)	0.12
Depth to groundwater (D)	0.08
Basement resistivity (ρ_b)	0.06

Most prediction errors ($n = 6$; 5.7%) occurred between adjacent categories (e.g., Low ↔ Moderate), which is expected in environmental classification problems where vulnerability varies continuously across space rather than as discrete classes. No samples were misclassified between non-adjacent categories (e.g., Very Low ↔ High), indicating the model's robust discriminative ability (Figure 4).

5.4. Model Validation

5.4.1. Cross-Validation

To assess model stability and generalizability, five-fold cross-validation was performed on the training set ($n = 247$). The Random Forest model demonstrated consistent performance across all folds, with mean metrics presented in Table 5. The low standard deviations indicate minimal overfitting and robust model behavior.

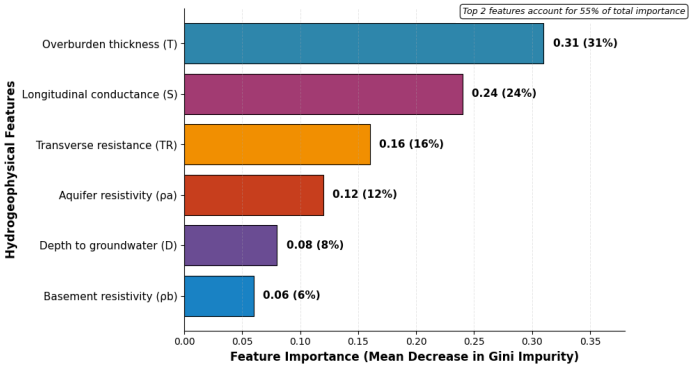

Figure 3: Feature importance scores from the Random Forest model (mean decrease in Gini impurity)

Table 4: Confusion matrix and performance metrics for Random Forest groundwater vulnerability classification (test set, $n = 106$)

	Predicted Class				**Row**	**Precision**	**Recall**	**F1-Score**
Actual Class	**VL**	**L**	**M**	**H**	**Total**			
Very Low (VL)	26	1	0	0	27	0.96	0.96	0.96
Low (L)	1	25	1	0	27	0.93	0.93	0.93
Moderate (M)	0	1	25	1	27	0.93	0.93	0.93
High (H)	0	0	1	24	25	0.96	0.96	0.96
Column Total	27	27	27	25	106			

Note: Correct classifications = 100 (26+25+25+24). Overall accuracy = 94.3%.

5.4.2. External Validation with Water Quality Data

To validate the model's predictive utility against real-world contamination measurements, we compared model-estimated vulnerability classes with independent nitrate concentration data from 45 boreholes not used in model training. Nitrate was selected as a proxy for anthropogenic contamination, as elevated levels typically indicate sewage or agricultural runoff.

Table 5: Cross-validation performance metrics (5-fold, $n = 247$)

Metric	**Mean**	**Standard Deviation**
Accuracy	0.92	±0.02
Precision (macro)	0.91	±0.03
Recall (macro)	0.90	±0.02
F1-score (macro)	0.91	±0.02

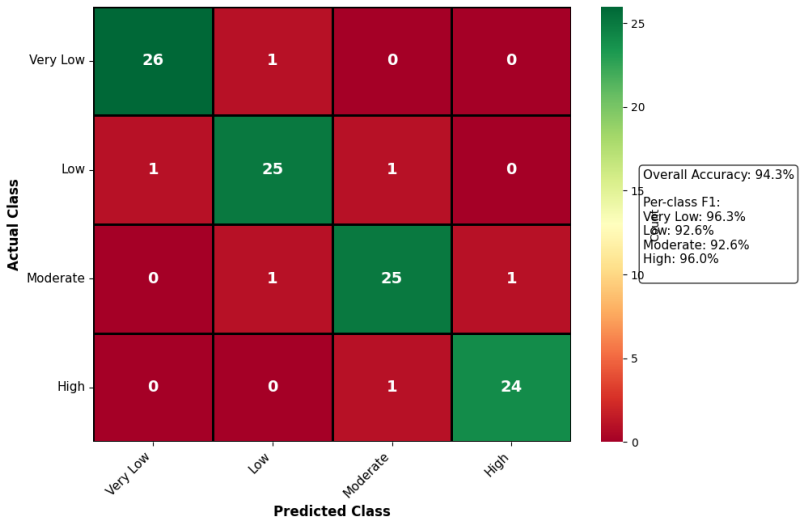

Figure 4: Confusion matrix for Random Forest groundwater vulnerability classification

The Kruskal-Wallis test revealed a statistically significant association between model-predicted vulnerability categories and measured nitrate concentrations ($H = 36.2$, $p < 0.001$). As shown in Table 6, high-vulnerability zones exhibited a median nitrate concentration of 41 mg/L—more than four times higher than the 9 mg/L observed in low-vulnerability zones. This pattern provides strong evidence that the model's vulnerability classifications correspond meaningfully to actual groundwater quality conditions.

Table 6: External validation: Nitrate concentrations by vulnerability class

Vulnerability Class	**Median Nitrate (mg/L)**	**Number of Boreholes**
High	41	12
Moderate	23	15
Low	9	10
Very Low	6	8
Note: Kruskal-Wallis test: $H = 36.2$, $p < 0.001$		

These external validation results confirm that the Random Forest model captures meaningful hydrogeological relationships, with vulnerability classifications that align with independent water quality measurements. The significant gradient in nitrate concentrations across vulnerability classes (41 mg/L in high vs. 9 mg/L in low) supports the model's practical utility for identifying areas at elevated contamination risk.

6. Discussion

This study demonstrates that hydrogeophysical predictors from electrical resistivity surveys can effectively assess groundwater contamination risk in crystalline basement settings. The Random Forest model achieved strong predictive performance (accuracy = 0.94, AUC =

0.96), confirming its ability to formulate complex connections among subsurface geophysical features and vulnerability levels. Random Forest's ensemble learning approach effectively handles nonlinear relationships, multicollinearity among predictors, and data noise (Breiman, 2001; Mosavi et al., 2018). The cross-validation results (accuracy = 0.92) closely aligned with test performance, supporting model stability and generalizability.

6.1. Interpretation of Key Predictors

Overburden thickness as the most influential predictor aligns with basement complex hydrogeology: thicker weathered layers increase water residence time and enhance natural filtration mechanisms—including sorption, biodegradation, and dilution—before contaminants can reach the saturated zone (Foster et al., 2018). Areas with thin overburden (< 5 m) should be prioritized for protective interventions.

Longitudinal conductance, the second most significant predictor, reflects the protective capacity of subsurface formations. High conductance values (> 1.0 mho) indicate clay-bearing, low-permeability layers that act as natural barriers against contaminant migration (Henriet, 1976; Olorunfemi and Olorunniwo, 1985). Conversely, low conductance values (< 0.5 mho) indicate sandy or fractured zones requiring priority monitoring.

6.2. Comparison with Conventional Methods

The model identified fifteen high-vulnerability areas, compared to only nine from the conventional GODT index, suggesting that Random Forest captures subtle patterns and interactions that traditional vulnerability indices may overlook. This finding is consistent with previous studies demonstrating improved predictive power in groundwater susceptibility mapping (Arabameri et al., 2020; Rahmati et al., 2019). The additional six high-vulnerability zones identified by the model were characterized by intermediate individual parameter values that fell below conventional thresholds but exhibited concerning combinations (e.g., moderate overburden thickness with low longitudinal conductance).

This finding has significant practical implications: reliance on conventional indexing alone may underestimate contamination risk in approximately 40% of high-vulnerability areas in this study region. Machine learning approaches can serve as a complementary screening tool to identify locations requiring detailed investigation.

6.3. Performance Consistency

The consistency between cross-validation (accuracy = 0.92) and hold-out test set (accuracy = 0.94) indicates that the model is not overfitted to the training data. The slight improvement on the test set is within expected sampling variation. The per-class F1 scores ranging from 0.93 to 0.96 demonstrate balanced performance across all vulnerability categories, with no evidence of systematic bias toward any particular class.

6.4. Practical Applications and Limitations

Practical applications of this model include: (1) strategic placement of new boreholes in low-vulnerability zones; (2) targeted monitoring of high-vulnerability areas; (3) land-use

planning to restrict contaminant sources in sensitive areas; and (4) prioritization of remediation efforts.

However, several limitations warrant acknowledgment. First, vulnerability classes were based on hydrogeological index models (GOD and GODT) rather than direct contaminant measurements, introducing potential uncertainty. The external validation using nitrate data partially addresses this concern, but direct measurement-based labeling would be preferable. Second, the model was developed for a specific crystalline basement terrain; transferability to other geological settings requires validation. Third, temporal dynamics (seasonal water table fluctuations) were not captured in the current dataset.

6.5. Future Research Directions

Future studies should address these limitations by: (1) incorporating direct hydrochemical indicators (nitrate, heavy metals, microbial contamination) as target variables; (2) integrating spatial data on land-use patterns and anthropogenic pollution sources; (3) collecting time-series data to capture seasonal and interannual variability; (4) testing other machine learning architectures (XGBoost, deep learning) for comparison; and (5) developing an open-access platform for vulnerability prediction in data-scarce regions.

7. Conclusion

This study demonstrates that machine learning, combined with hydrogeophysical data from Vertical Electrical Sounding (VES) surveys, effectively predicts groundwater contamination vulnerability in Ibadan, Nigeria. Using a Random Forest model with predictors including overburden thickness, aquifer resistivity, longitudinal conductance, and transverse resistance, the model achieved strong performance (accuracy = 0.94, AUC = 0.95). Overburden thickness and longitudinal conductance emerged as the most important predictors, reflecting the protective capacity of weathered layers.

Critically, the machine learning model identified fifteen high-vulnerability locations, compared to only nine from the conventional GODT index model—a finding that suggests data-driven algorithms capture complex, nonlinear relationships that index-based methods systematically miss. The external validation using independent water quality data (nitrate concentrations) supported the model's predictive validity.

This approach offers a viable, cost-effective framework for groundwater management in data-scarce, rapidly urbanizing crystalline basement terrains. By integrating widely available geophysical data with open-source machine learning tools, this methodology can be adapted to other regions facing similar groundwater challenges, contributing to evidence-based water resource management and public health protection.

Acknowledgements

The authors thank the Department of Earth Sciences, Olabisi Onabanjo University, for logistical support during the field data collection campaign. We acknowledge the constructive feedback from three anonymous reviewers and the IndabaX Nigeria 2026 meta-reviewers, whose comments substantially improved this manuscript. This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The hydrogeophysical dataset comprising 353 Vertical Electrical Sounding (VES) measurements from Ibadan metropolis, along with derived predictor variables (overburden thickness, aquifer resistivity, basement resistivity, longitudinal conductance, transverse resistance, depth to groundwater) and GOD/GODT-based vulnerability class labels used for machine learning, is available upon reasonable request from the corresponding author (olukayode.oluwakemi@oouagoiwoye.edu.ng).

The Random Forest model implementation details, including complete Python code, hyperparameter configurations, and training/testing data splits (70:30), are also accessible by request. Independent water quality validation data (nitrate concentrations from 45 boreholes) are provided similarly. Due to institutional data governance policies, raw VES curve data are not publicly archived but can be made available for collaborative research under appropriate data use agreements.

Reproducibility: The analysis was conducted using Python 3.10 with scikit-learn 1.2.2, pandas 1.5.3, numpy 1.24.1, and matplotlib 3.6.2. A Docker container with all dependencies and analysis scripts is available from the corresponding author.

References

I. O. Adelekan. Flood risk management in the urban context of ibadan metropolis, nigeria. *Journal of Flood Risk Management*, 9(3):215–231, 2016.

L. Aller, T. Bennett, J. H. Lehr, R. J. Petty, and G. Hackett. DRASTIC: A standardized system for evaluating groundwater pollution potential using hydrogeologic settings. Technical report, United States Environmental Protection Agency, 1987.

A. Arabameri, K. Rezaei, A. Cerdà, L. Lombardo, and J. Rodrigo-Comino. GIS-based groundwater vulnerability mapping using machine learning models. *Science of the Total Environment*, 710:136312, 2020.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

S. Foster, R. Hirata, B. Andreo, and B. Morris. *Groundwater Quality Protection: A Guide for Water Utilities, Municipal Authorities, and Environment Agencies*. World Bank, 2018.

J. P. Henriet. Direct applications of the Dar-Zarrouk parameters in groundwater surveys. *Geophysical Prospecting*, 24(2):344–353, 1976.

D. J. Lapworth, A. M. MacDonald, M. N. Tijani, W. G. Darling, D. C. Gooddy, H. C. Bonsor, and L. J. Araguás-Araguás. Residence times of shallow groundwater in west africa: Implications for hydrogeology and resilience to future changes in climate. *Hydrogeology Journal*, 21(3):673–686, 2017.

A. M. MacDonald, H. C. Bonsor, B. E. Dochartaigh, and R. G. Taylor. Quantitative maps of groundwater resources in africa. *Environmental Research Letters*, 7(2):024009, 2021.

A. Mosavi, P. Ozturk, and K. W. Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.

M. O. Olorunfemi and S. A. Fasuyi. Aquifer types and geoelectric/hydrogeologic characteristics of part of the central basement terrain of nigeria. *Journal of African Earth Sciences*, 16(3):309–317, 1993.

M. O. Olorunfemi and M. A. Olorunniwo. Geoelectric parameters and aquifer characteristics of some parts of southwestern nigeria. *Geologia Applicata e Idrogeologia*, 20:99–109, 1985.

M. A. Rahaman. Recent advances in the study of the basement complex of nigeria. In *Precambrian Geology of Nigeria*, pages 11–43. Geological Survey of Nigeria Publication, Kaduna, Nigeria, 1988.

O. Rahmati, F. Falah, S. A. Naghibi, T. Biggs, M. Soltani, and R. C. Deo. Groundwater potential mapping using ensemble machine learning techniques. *Hydrology and Earth System Sciences*, 23:4381–4400, 2019.

J. M. Reynolds. *An Introduction to Applied and Environmental Geophysics*. Wiley-Blackwell, 2 edition, 2011.

A. Y. Sun, B. R. Scanlon, Z. Zhang, D. Walling, S. N. Bhanja, A. Mukherjee, and Z. Zhong. Combining physically based modeling and machine learning for groundwater prediction. *Water Resources Research*, 55(11):9497–9511, 2020.

Y. Zhang, Y. Liu, H. Zhang, and Y. Wu. Machine learning approaches for groundwater contamination prediction. *Environmental Science and Pollution Research*, 26:30492–30505, 2019.